UNIVERSIDAD DE GRANADA

Programa de Doctorado en Estadística Matemática y Aplicada

Tesis Doctoral

Aportaciones sobre el uso de información auxiliar para la estimación de medidas de desigualdad y pobreza.

María Dolores Illescas Manzano

Tesis dirigida por

Prof. María del Mar Rueda García

Prof. Sergio Martínez Puertas

Granada, Enero, 2023

Editor: Universidad de Granada. Tesis Doctorales
Autor: María Dolores Illescas Manzano
ISBN: 978-84-1117-862-4
URI: https://hdl.handle.net/10481/82097

Índice general

Índice de tablas

Dedicatoria

Dedico esta tesis a mi tía María José y mi padre Enrique porque desde el cielo tienen que estar muy orgullosos de su niña.

A mi madre, Loli y mis suegros Alfredo y Adela por el apoyo continuo e incondicional; a mis hijos María José y Sergio por la paciencia y entendimiento del tiempo dedicado a la elaboración de mi tesis.

A mi directora de tesis, María del Mar por su sabiduría y conocimientos aportados durante mi periodo doctoral.

A mi codirector, Sergio por el tiempo dedicado y ayuda incondicional para la consecución de las metas propuestas.

Agradecimientos

Gracias de todo corazón a mis directores de tesis, los doctores María del Mar Rueda García, Catedrática de la Universidad de Granada y Sergio Martínez Puertas, Profesor Titular de Universidad de Almería. Gracias por vuestra dedicación, ánimo, perseverancia, conocimiento y trabajo. Ha sido un honor en mayúsculas poder disponer de vuestra dirección y experiencia para saber en cada momento como dirigir y orientarme en la consecución de mi tesis doctoral.

No se me puede olvidar mencionar a mi madre, su apoyo incondicional durante estos años ha sido crucial para que mis hijos María José y Sergio se hayan sentido arropados, atendidos, queridos y yo poder dedicar mi tiempo a la elaboración de mi tesis doctoral. Siempre agradecida.

A lo más importante de mi vida, mis hijos. Sergio y María José, gracias por tener siempre una sonrisa y haber aprendido a tener paciencia mientras mamá tenía que trabajar para conseguir con éxito su tesis doctoral.

Y por último a mi marido Sergio, sin él, esta tesis doctoral hubiera sido más difícil realizarla. Sin su dedicación, perseverancia, estudio, investigación y capacidad de trabajo no hubiera podido culminar con éxito las diferentes publicaciones que integran mi tesis doctoral.

Resumen

La función de distribución es un parámetro funcional y no lineal a partir del cual se pueden estimar otros parámetros poblacionales de interés como los cuantiles, la función de fiabilidad o el índice de Gini y por ello la función de distribución es especialmente atractiva para diversas áreas de investigación. Así, medidas como los cuantiles poblacionales pueden ser parámetros de gran interés por diversas áreas tanto pertenencientes a la investigación de Ciencias Sociales como a la investigación de Ciencias y que abarca disciplinas tan diversas como antropología, física, pediatría, edafología, sociología, psicología o economía. En este último campo, los cuantiles alcanza una gran relevancia dado el especial interés que los gobiernos, organismos oficiales y la investigación económica ha mostrado en el análisis de la pobreza y desigualdad salarial pues muchas de las medidas empleadas en estos estudios están basadas en el cálculo de cuantiles o ratios de los mismos.
Por otro lado, la información obtenida por organismos oficiales y por agencias estadísticas a la hora de medir la pobreza no sólo se restringe a información relativa al nivel salarial, sino que abarca aspectos relacionados con las condiciones de vida y variables sociodemográficas. De este modo, en los estudios de pobreza se dispone de un volumen relevante de información auxiliar que si se incorpora en la fase de estimación, puede mejorar la precisión y fiabilidad de las mediciones de pobreza a través de cuantiles. Por ello, es esencial disponer de técnicas de estimación para la función de distribución que incorporen la información auxiliar disponible a la hora de estimar la función de distribución.

Dada la relevancia de la función de distribución, en la investigación sobre muestreo en poblaciones finitas podemos encontrar una gran variedad de estimadores indirectos de la función de distribución, esto es, estimadores que incorporan información auxiliar disponible para mejorar la eficiencia del estimador. Si bien, una de las cuestiones esenciales a la hora de desarrollar estimadores de la función de distribución, es que el estimador obtenido respete las propiedades de la función de distribución y muchas de estas alternativas no satisfacen todas las propiedades, no pudiendo emplearse en la estimación de cuantiles. Por ello, el principal propósito de esta tesis es la obtención de estimadores para la función de distribución que

incorporen información auxiliar y que sean auténticas funciones de distribución, para así emplearlos en la estimación de cuantiles y medidas de pobreza.

Recientemente, el método de calibración originalmente propuesto por Deville & Särndal (1992) para la estimación de totales y medias ha sido empleado en la formulación de nuevos estimadores indirectos de la función de distribución. Una de las propuestas que destaca por su simplicidad computacional es el estimador propuesto por Rueda et al. (2007a). Otra ventaja que presenta este estimador es que es satisface todas las propiedades de función de distribución bajo restricciones leves. Sin embargo, su comportamiento asintótico puede verse afectado por el conjunto de puntos auxiliares con los que se lleva a cabo el proceso de calibración y si bien la elección óptima del vector de puntos auxiliares ha sido determinada (Martínez et al., 2017), no se ha analizado si el estimador basado en la elección óptima satisface las propiedades de función de distribución. En la tesis se analizarán las condiciones bajo las cuales este estimador es una genuina función de distribución y se formularán los correspondientes estimadores de cuantiles y ratios de percentiles asociados al mismo

Adicionalmente, el estimador de Rueda et al. (2007a), al ser un estimador calibrado, puede sufrir sobre-calibración (Chauvet & Goga, 2022) cuando la dimensión del vector de puntos auxiliares es elevada, lo que sucede frecuentemente con la elección óptima propuesta en Martínez et al. (2017). Para resolver esta cuestión, en la tesis se analizará teóricamente las condiciones bajos las cuales se puede reducir la dimensión óptima del vector propuesto en Martínez et al. (2017) sin que la eficiencia empeore.

Finalmente, dado que los estudios de pobreza, desigualdad y condiciones de vida tratan cuestiones sensibles y son principalmente desarrollados mediante encuestas, es inevitable la presencia de datos faltantes por diversas razones (ausencia del encuestado en el momento de la encuesta, negativa a responder sobre el nivel de ingresos, etc) y en consecuencia, es necesario disponer de técnicas de estimación de la función de distribución que permitan tratar la falta de respuesta. La presente tesis formulará mediante el uso del método de calibración, estimadores de la función de distribución diseñados para tener en cuenta la falta de respuesta.

Esta tesis se presenta como un compendio de cuatro publicaciones directamente vinculadas con los contenidos y probelmas de iinvestigación tratados en la presente tesis. Entre las cuatro publicaciones, tres de ellas son artículos publicados en revistas indexadas en el Journal Citation Reports del Science Citation Index y correspondientes al primer cuartil todas ellas, mientras que la otra aportación se trata de un capítulo de libro que actualmente también se encuentra publicado en una editorial indexada en el índice SPI Scholarly Publishers Indicators y que se encuentra en la cuarta posición del ranking general. En la segunda parte de la tesis se presentan las versiones completas de las cuatro publicaciones ordenadas en Apéndices por

orden cronológico. Previamente, en los capítulos presentados en la primera parte de la tesis se presentan los aspectos fundamentales para el desarrollo de las publicaciones y que permiten facilitar la lectura de la tesis. Así, en el primer capítulo se introduce la importancia de la función de distribución, cuantiles y medidas de pobreza y se exponen problemas abiertos en este campo. En el segundo capítulo se formulan los objetivos que se pretenden alcanzar con la presente tesis. En el capítulo 3 se expone la metodología empleada en el desarrollo de la tesis. Los capítulos 4 y 5 se exponen respectivamente los resultados y conclusiones de mayor relevancia alcanzados en la tesis. Finalmente, el capítulo 6 está dedicado a las líneas de investigación actuales en las que se está ya trabajando y futuras líneas de investigación.

Parte I

Capítulo 1

Introducción

La estimación de la función de distribución es un tema de estudio importante en la investigación mediante encuestas y ha recibido mucha atención en los últimos años, debido a que presenta características valiosas para la aplicación en muchas áreas. Así, la función de distribución permite, por ejemplo, obtener la función de fiabilidad, empleada en los análisis de datos de vida y en fiabilidad de sistemas (Acal et al., 2019). También a través de la función de distribución se puede contrastar si dos muestras provienen de una misma población (Alba-Fernández et al., 2017).

Adicionalmente, la función de distribución permite obtener medidas importantes como los cuantiles poblacionales, de modo que si disponemos de estimadores de la función de distribución que satisfacen las propiedades de función de distribución, podemos emplearlos en la estimación de cuantiles, cuestión de gran interés en diversas áreas de investigación tales como antropología (Bogin & Sullivan, 1986); ciencias de la salud (Bohn et al., 2019; Kimbro et al., 2011; Tellez-Plaza et al., 2008) y pediatría (Vander Wal & Mitchell, 2011), toxicología (Wolford et al., 1986), procesos químicos y físicos (Wilson et al., 2012), edafología (Bu et al., 2015), sociología (Eisenberg et al., 2005; Kimbro et al., 2011), psicología (Crawford & Garthwaite, 2009) o economía (Decker et al., 2014; Gelman et al., 2010) en las cuales algunas medidas e indicadores dependen de cuantiles.

Específicamente, en el área de economía, la medición de pobreza y la desigualdad salarial así como los estudios sobre exclusión social son temas de gran interés tanto para la investigación económica (Darvas, 2019; Meyer & Sullivan, 2012; Sompolska-Rzechu la & Kurdyś-Kujawska, 2022) como para los gobiernos, instituciones oficiales y sociedad en general (European Commission, 2010a; Eurostat Statistics, 2020; Jones & Weinberg, 2000; Meglio, 2018), pues medidas como la tasa de pobreza oficial y el número de personas en situación de pobreza son importantes para determinar el grado de bienestar económico de un país. Así, la Comisión Europea en su comunicación del año 2010 fijó la estrategia Europa 2020 y en ella uno de

los objetivos consiste en reducir en un 25 % el número de europeos que viven por debajo de los umbrales nacionales de pobreza, lo que supone rescatar a 20 millones de europeos de exclusión social (European Commission, 2010a). Para ello, se fijó como una de las iniciativas la creación de la "Plataforma europea contra la pobreza", cuyo objetivo es garantizar la cohesión social y fomentar condiciones idóneas para que personas en riesgo de exclusión social puedan vivir en condiciones dignas (European Commission, 2010b).

Dado que en los estudios de pobreza están basados en variables como los salarios o ingresos, que suelen tener distribuciones asimétricas, los cuantiles son medidas más adecuadas que por ejemplo la media y por ello algunos índices y medida de pobreza están basados en cuantiles. Así, por ejemplo, entre las medidas de pobreza incluidas en los análisis de pobreza y condiciones de vida que están basadas en la función de distribución y en cuantiles, podemos mencionar la línea o umbral de pobreza, definida como el umbral de ingresos por debajo del cual un individuo es considerado en situación de pobreza y que Eurostat fija como el 60 por ciento de la mediana del ingreso neto equivalente (Eurostat Statistics, 2022). También podemos mencionar el índice de de recuento de pobreza o tasa de pobreza, que se define como la proporción de individuos que están en riesgo de pobreza, esto es, personas con una renta disponible por debajo del umbral de pobreza y que puede ser estimado mediante la función de distribución (Martínez et al., 2020).

Por otro lado, la desigualdad salarial también es un aspecto clave a la hora de analizar y comprender cómo se genera la pobreza (Guio et al., 2021) siendo una prioridad de la Comisión Europea el desarrollo de indicadores tanto para medir la pobreza como la desigualdad salarial (Eurostat Experimental statistics, 2022). Algunas de las medidas empleadas para evaluar la desigualdad salarial están basadas en ratios de quantiles, así, Eurostat para analizar la desigualdad salarial en la Unión Europea emplea el ratio de percentiles P80/P20 (Eurostat Products Datasets, 2022) mientras que el "European Trade European Union Institute" emplea el ratio de percentiles P90/P10 (Countouris et al., 2020) al igual que el "US Census Bureau" que adicionalmente para el análisis de la desigualda salarial en EEUU también considera los ratios P95/P20; P95/P50; P80/P50; P80/P20 y P20/P50 (Shrider et al., 2021). De igual forma, la investigación económica previa ha considerado los ratios de percentiles como indicadores para medir la desigualdad salarial, como los ratios P95/P50 (Machin et al., 2003), los ratios P90/P10; P95/P20 y P80/P20 (Jones & Weinberg, 2000); los ratios P50/P5 y P50/P25 (Dickens & Manning, 2004) y el ratio P50/P10 (Burtless, 1999). Dado que la estimación de cuantiles está estrechamente relacionada con el análisis y estudio de pobreza, la disponibilidad de técnicas que permitan estimar la función de distribución y los cuantiles se ha vuelto crucial en este tipo de estudios.

Por otro lado, la información obtenida a partir de las encuestas realizadas por organismos oficiales y por agencias estadísticas cuyo objetivo es la medición de la pobreza, además de proporcionar información acerca de la variable de estudio, esto es, nivel de ingresos de las personas y hogares, también proporcionan información sobre variables adicionales relacionadas con las condiciones y características de vida de los

individuos, y por tanto es habitual disponer de información relativa a edad, sexo, educación, empleo, etc (INE, 2022). Estas variables proporcionan información auxiliar relacionada con la variable de estudio, que puede ser incorporada en la fase de estimación para mejorar la precisión y eficiencia de un estimador, dando lugar así, a los denominados estimadores indirectos. Dado que en los análisis de pobreza se dispone de un gran volumen de información auxiliar relevante, es fundamental disponer de técnicas de estimación indirectas para la función de distribución, es decir, estimadores que incorporen y hagan un uso eficiente de la información auxiliar disponible a la hora de estimar la función de distribución.

En la literatura previa, existe una amplia variedad de técnicas indirectas para la estimación de la función de distribución. Así, entre los estimadores indirectos para la función de distribución basados en el diseño, podemos mencionar entre otros, los estimadores de razón y diferencia (Rao et al., 1990), el estimador de Kuk y Mak (Kuk & Mak, 1989), o el estimador Silva y Skinner (Silva & Skinner, 1995). Entre las alternativas existentes basadas en el modelo, tenemos el estimador de Chambers y Dunstan (Chambers & Dunstan, 1986), el estimador de Rao, Kovar y Mantel (Rao et al., 1990) o el estimador de Wang y Dorfman (Wang & Dorfman, 1996).

Si bien existe una amplia gama de procedimientos que permiten la incorporación de la información auxiliar disponible en la estimación de la función de distribución, algunas de estas alternativas presentan inconvenientes. Así, los estimadores de razón, diferencia y el estimador de Rao, Kovar y Mantel no son funciones de distribución genuinas, lo que dificulta su empleo en la estimación de cuantiles poblacionales y en las medidas de pobreza anteriormente comentadas. Por otro lado el estimador de Silva y Skinner depende de la elección de los estratos y no proporciona estimaciones perfectas para las variables auxiliares de forma general. Finalmente, los estimadores basados en el modelo depende de la elección del mismo y requieren un coste computacional considerable.

Por otro lado, a la hora de incorporar información auxiliar, el método de calibración (Deville & Särndal, 1992), originalmente empleado como método de reponderación en la estimación de totales poblacionales, es considerado como un importante instrumento metodológico en la producción de estadísticas a gran escala por parte de varias agencias nacionales de estadística (Lafferty & McCormack, 2015; Le Guennec & Sautory, 2002; Memobust, 2014), de manera que han desarrollado software diseñado que permite calibrar la información auxiliar disponible en registros administrativos y otras fuentes (Le Guennec & Sautory, 2002; Vanderhoeft, 2001).

Los estudios previos han considerado diferentes implementaciones del método de calibración para el desarrollo de nuevos estimadores de la función de distribución y cuantiles (Harms & Duchesne, 2006; Rueda et al., 2007a; Wu, 2003). Asumiendo un modelo de superpoblación general Wu (2003) propuso un estimador basado en el modelo que es óptimo para la esperanza bajo el modelo de la varianza. Entre los

incovenientes de la propuesta de Wu (2003) podemos mencionar, que en general el estimador propuesto no satisface las propiedades de función de distribución en general y requiere la estimación de parámetros del modelo de superpoblación asumido, ya que suelen depender de la variable de estudio lo que puede restringir su aplicación. Finalmente, para garantizar el comportamiento asintótico del estimador, se requieren restricciones adicionales sobre el diseño muestral (Chen & Wu, 2002).

Recientemente Rueda et al. (2007a), desarrollaron una familia de estimadores calibrados para la función de distribución, que es simple desde el punto de vista computacional y bajo condiciones suaves los estimadores obtenidos son funciones de distribución genuinas y ello permite su aplicación en la estimación de cuantiles y medidas de pobreza (Martínez Puertas & Martínez Puertas, 2013; Rueda et al., 2007b).

Ahora bien, uno de los inconvenientes que presenta el estimador propuesto por Rueda et al. (2007a) es que su comportamiento asintótico y su eficiencia depende de un vector de puntos auxiliares empleados en el proceso de calibración. Si bien el vector de puntos óptimo y la dimensión óptima del vector que hay que emplear en el proceso de calibración para optimizar la eficiencia en el estimador de Rueda et al. (2007a) ha sido determinado bajo muestreo aleatorio simple (Martínez et al., 2017), el problema que presenta este estimador, es que en general, el vector óptimo depende del punto t de la variable de estudio en el que se desea estimar la función de distribución, lo que no ocurría en el estimador de Rueda et al. (2007a) e implica que las reponderaciones empleadas en la calibración dependan del valor de t. Todo ello podría implicar que el estimador obtenido no respete las propiedades de función de distribución, lo que dificulta su aplicación en estimación de cuantiles y medidas de pobreza.

Adicionalmente, en muchas ocasiones el vector de puntos auxiliares que optimiza la eficiencia tiene una dimensión elevada lo que dificulta el proceso de calibración. Recientemente, se ha constatado que llevar a cabo el proceso de calibración cuando la dimensión de la información auxiliar es muy elevada puede provocar diversos problemas. En primer lugar, si se emplea un número elevado de variables auxiliares puede producirse sobrecalibración que puede provocar que el sesgo del estimador calibrado no sea despreciable en comparación con la varianza (Nascimento Silva & Skinner, 1997) y también puede empeorar la eficiencia del estimador (Chauvet & Goga, 2022). Finalmente, en el caso de la estimación de la función de distribución, una dimensión elevada de la información auxiliar puede provocar restricciones de calibración incompatibles.

Por otro lado, dado que los estudios de pobreza, desigualdad y condiciones de vida son principalmente desarrollados mediante encuestas y estudios basados en muestras, es habitual que se produzca la presencia de datos faltantes por diversas razones, tales como la no presencia del encuestado en el momento de la encuesta o porque el encuestado puede negarse a responder sobre su nivel de ingresos dado que el objetivo de los estudios de pobreza constituye una cuestión problemática. La presencia de datos faltantes y la reducción del tamaño de la muestra asociada a la misma puede provocar sesgos en las estimaciones y un aumento de la

varianza del estimador si los datos faltantes siguen algún patrón.

Aunque existen diversos métodos cuyo objetivo es obtener un conjunto de datos completo en las etapas de recopilación y procesamiento de datos, es posible enfrentar errores y pérdidas de entradas incluso después de que los datos hayan sido recopilados y filtrados y por tanto en ocasiones no es posible evitar la presencia de datos faltantes en la etapa de estimación. El tratamiento de falta de respuesta en la fase de estimación puede ser enfocado desde dos alternativas principalmente, la imputación y la reponderación.

En relación a la última alternativa, el objetivo consiste en la estimación mediante un conjunto de ponderaciones, basadas en la información auxiliar disponible, para las unidades muestrales que no presentan falta de respuesta. De este modo, el método de calibración, concebido originalmente para corregir el error de muestreo, ha sido aplicado para el ajuste del sesgo producido por la falta de respuesta en la estimación de parámetros como el total, media o proporciones (Andersson & Särndal, 2016; Kott & Liao, 2017; Särndal & Lundström, 2005) y existe disponible una extensa literatura dedicada a la estimación de la media poblacional en presencia de datos faltantes (Beaumont, 2005; Chang & Kott, 2008; Deville, 2000; Kott & Liao, 2012, 2015, 2017; Lesage et al., 2019; Jo et al., 2015). En particular, las propuestas de Deville (2000) y Kott & Liao (2017) se basan en calibración con variables instrumentales, donde se ajusta por medio de calibración tanto la falta de respuesta como el error de muestreo. Así, por un lado tenemos un conjunto de variables auxiliares empleadas en la modelización de la falta de respuesta mientras que se empleará otro vector de variables auxiliares para el proceso de calibración, llamadas variables instrumentales. Si bien, algunas de las variables incluidas en el modelo de falta de respuesta pueden también ser empleadas como variables instrumentales en el proceso de calibración, en la propuesta de Deville (2000) la dimensión del vector empleado en la modelización de falta de respuesta y del vector de variables instrumentales debe coincidir. Recientemente, Kott & Liao (2017) extendieron la propuesta de Deville (2000) al caso donde existen más variables instrumentales o de calibración que variables para el modelo de falta de respuesta.

Por el contrario, el desarrollo de técnicas de estimación de la función de distribución que incorporen el uso de información auxiliar en el tratamiento de la falta de respuesta es menos habitual, ni tampoco es extenso el uso del método de calibración para el tratamiento de la falta de respuesta en la estimación de la función de distribución, siendo necesario adaptar las diferentes metodologías existentes para la estimación de totales y medias, de forma que al aplicarlas en la estimación de la función de distribución, se obtengan estimadores que respeten las propiedades de función de distribución y así poder estimar cuantiles y medidas de pobreza. Todo ello, hace que la estimación de cuantiles y medidas de pobreza en presencia de falta de respuesta no haya sido tan extensamente tratada como otros parámetros poblacionales.

Por todo lo anteriormente mencionado, el principal propósito que nos planteamos en esta tesis doctoral es desarrollar estimadores de la función de distribución mediante técnicas de calibración que sean genuinas

funciones de distribución para así poder emplearlos en la estimación de cuantiles y medidas de pobreza. Con este propósito, en Apéndice 1, vamos a abordar la estimación de la función de distribución en presencia de falta de respuesta de forma que se procurará analizar bajo qué condiciones los estimadores obtenidos satisfacen todas las propiedades de función de distribución. Con ello, pretendemos obtener estimadores para cuantiles y medidas de pobreza cuando se produce falta de respuesta. Así, en el Apéndice 2 abordaremos esta cuestión y se llevará a cabo un estudio de simulación con datos reales procedentes de la Encuesta de condiciones de vida realizada por el Instituto Nacional de Estadística (INE) en España (INE, 2015). En relación también a la estimación de cuantiles y medidas de pobreza, en el Apéndice 3, trataremos de analizar bajo qué condiciones la metodología propuesta por Martínez et al. (2017) puede ser aplicada en la estimación de cuantiles y medidas de pobreza. Finalmente, en el Apéndice 4, abordaremos el problema de la alta dimensionalidad que en muchas ocasiones presenta el vector óptimo de calibración en el que se basa la propuesta de de Martínez et al. (2017) y evitar así los problemas derivados de esta alta dimensionalidad.

Capítulo 2

Objetivos

Como se ha puesto de manifiesto en la Introducción, a pesar de que la investigación previa se ha interesado en la estimación indirecta de la función de distribución, existen todavía algunas cuestiones que requieren de una mayor investigación y que son la principal motivación para el desarrollo de esta tesis doctoral. Por ello, el principal propósito de la presente tesis es investigar más a fondo algunos temas escasamente abordados hasta ahora en relación a la estimación de la función de distribución y cuantiles. Este propósito general se hará más concreto en este capítulo a través de la definición de objetivos generales y específicos.

Entre los objetivos generales que nos proponemos alcanzar en esta tesis doctoral podemos mencionar:

1. Desarrollar estimadores de la función de distribución mediante técnicas de calibración que sean genuinas funciones de distribución para así poder emplearlos en la estimación de cuantiles y medidas de pobreza.

2. Desarrollar técnicas de estimación de la función de distribución en presencia de falta de respuesta capaces de reducir el sesgo asociado a esta falta de respuesta.

3. Desarrollar técnicas de calibración que eviten la alta dimensionalidad en las restricciones de calibración.

De igual forma, con la presente tesis doctoral nos proponemos alcanzar los siguientes objetivos específicos:

1. Desarrollo de técnicas de estimación para la función de distribución bajo falta de respuesta mediante el empleo del método de calibración.

2. Extender la técnica de calibración para el tratamiento falta de respuesta para la media poblacional propuestas en Kott & Liao (2017) para la estimación de la función de distribución y analizar bajo qué condiciones los estimadores obtenidos son auténticas funciones de distribución.

3. Adaptación de la técnica de estimación propuesta por Rueda et al. (2007a) para la función de distribución en presencia de datos faltantes.

4. Aplicación de las técnicas desarrolladas en los objetivos específicos anteriores en la estimación de cuantiles y medidas de pobreza.

5. Analizar y establecer bajo qué condiciones el estimador calibrado de la función de distribución propuesto por Martínez et al. (2017) satisface las propiedades de función de distribución para posteriormente analizar su comportamiento a la hora de estimar cuantiles y medidas de pobreza.

6. Analizar teóricamente si es posible la reducción de la dimensión del vector óptimo de calibración obtenido en Martínez et al. (2017) sin pérdida de eficiencia.

En todos y cada uno de los objetivos, se tratará de obtener expresiones de la varianza asintótica de los estimadores propuestos pero dado que tanto los cuantiles como las medidas de pobreza no son parámetros lineales, en ocasiones obtener una fórmula teórica para su varianza no será posible. En los casos donde el comportamiento asintótico del estimador no pueda ser establecido de forma teórica, emplearemos técnicas bootstrap para la estimación de varianza.

2.1. Treating nonresponse in the estimation of the distribution function

En primer lugar, en el Apéndice 1, vamos a considerar la extensión de las técnicas propuestas por Deville (2000) y Kott & Liao (2017) para la estimación de función de distribución en presencia de falta de respuesta. De igual forma trataremos la adaptación de la metodología propuesta en Rueda et al. (2007a) para la estimación de la función de distribución bajo la presencia de falta de respuesta.
En todos los casos, se aborda el análisis de las condiciones bajo las cuales las técnicas propuestas satisfagan las propiedades de función de distribución Con ello, se cubren los objetivos generales 1 y 2 y los objetivos específicos 1,2 y 3.

2.2. Calibration adjustment for dealing with nonresponse in the estimation of poverty measures

Una vez desarrollado técnicas eficientes para la estimación de función de distribución que bajo condiciones poco restrictivas son genuinas funciones de distribución, en el Apéndice 2 se aborda su extensión a la estimación de cuantiles y medidas de pobreza en presencia de datos faltantes.
Dado que tanto los cuantiles y las medidas de pobreza (especialmente aquellas basadas en ratios) no son parámetros lineales, es necesario recurrir a técnicas bootstrap para la estimación de la varianza de los estimadores propuestos.
Finalmente, un estudio de simulación con datos reales procedente de la Encuesta de condiciones de vida realizada por el Instituto Nacional de Estadística (INE) en España (INE, 2015) es llevado a cabo para analizar el comportamiento de los estimadores propuestos frente a otras alternativas.
Con todo ello, se alcanza la consecución del objetivo general 1 y del objetivo específico 4.

2.3. The optimization problem of quantile and poverty measures estimation based on calibration

En tercer lugar, en el Apéndice 3, trataremos de analizar bajo qué condiciones el estimador propuesto por Martínez et al. (2017) es una auténtica función de distribución y así poder también aplicar esta metodología en la estimación de cuantiles y medidas de pobreza. Más concretamente, asumiendo que el diseño muestral empleado es muestreo aleatorio simple, se establece bajo condiciones leves que el estimador propuesto por Martínez et al. (2017) es monótono no decreciente y en consecuencia puede ser aplicado en la estimación de cuantiles y medidas de pobreza. De este modo, se aborda su aplicación a la estimación de dichos parámetros y se analiza el comportamiento de las técnicas propuestas con un estudio de simulación con datos reales nuevamente procedentes de la de la Encuesta de condiciones de vida (INE, 2015), siendo nuevamente necesario recurrir a técnicas bootstrap para la estimación de la varianza.
De este modo, el objetivo general 1 y el objetivo 5 son alcanzados.

2.4. Reduction of optimal calibration dimension with a new optimal auxiliary vector for calibrated estimators of the distribution function

Finalmente, en el Apéndice 4, abordaremos el problema de la alta dimensionalidad que en muchas ocasiones presenta el vector óptimo de calibración en el que se basa la propuesta de de Martínez et al. (2017)

y que puede generar sobrecalibración. Así, trataremos de analizar si es posible alcanzar la misma eficiencia pero empleando un vector de menor dimensión, lo que además de evitar la sobrecalibración tambien nos permitirá simplificar el coste computacional. A lo largo del Apéndice 4 se establecen bajo una generalidad de situaciones las condiciones bajo las cuales la dimensión del vector óptimo propuesto en Martínez et al. (2017) puede ser reducida teóricamente. Mediante estudios de simulación se evidencia que el coste computacional requerido con la nueva propuesta se ve sustancialmente reducido sin pérdida de eficiencia e incluso en ocasiones mejorando la eficiencia de la propuesta original de Martínez et al. (2017). Con ello, se cubre el objetivo general 3 y el objetivo específico 4.

Capítulo 3

Metodología

Para la consecución de los objetivos propuestos en esta tesis doctoral será necesario recurrir a una amplia variedad de técnicas indirectas de estimación de la función de distribución. Así, se revisarán aquellas técnicas de estimación indirectas de la función de distribución que serán incluidas en los estudios de simulación con própositos comparativos. Adicionalmente, dado que varios objetivos están relacionados con los estimadores de Rueda et al. (2007a) y Martínez et al. (2017) revisaremos en profundidad ambas técnicas de estimación. De igual forma será necesario considerar técnicas para la estimación de la varianza de los nuevos estimadores desarrollados.
También será necesario una revisión de las medidas de pobreza basadas en cuantiles así como la revisión de diferentes técnicas para el tratamiento de falta de respuesta para su adaptación a la estimación de la función de distribución.

3.1. Estimación de la función de distribución en poblaciones finitas

Si consideramos una población finita $U = \{1, 2, \ldots, N\}$ formada por N unidades diferentes y donde se ha definido un diseño muestral $p(\cdot)$ con probabilidades de inclusión de primer y segundo orden dadas respectivamente por $\pi_k > 0$ and $\pi_{kl} > 0$ $k, l \in U$. Consideremos una muestra $s = \{1, 2, \ldots, n\}$ de tamaño fijo n que es seleccionada de acuerdo al diseño muestral $p(\cdot)$ y denotaremos por $d_k = \pi_k^{-1}$ los pesos básicos del diseño muestral $p(\cdot)$, que asumiremos conocidos para todas las unidades poblacionales $k \in U$. Sea y_k el valor de la variable de estudio para la unidad poblacional k y sea $\mathbf{x}_k^{'} = (x_{1k}, \ldots, x_{Jk})$ un vector de variables auxiliares de dimensión J para la unidad poblacional k. Asumiremos que el valor $\mathbf{x}_k$ está disponible para todas las unidades poblacionales mientras que el valor y_k sólo está disponible para las unidades muestrales. El objetivo es la estimación de la función de distribución de la variable de estudio y, cuyo valor en un punto

t viene dado por:

$$F_y(t) = \frac{1}{N}\sum_{k\in U}\Delta(t-y_k) \tag{3.1}$$

donde

$$\Delta(t-y_k) = \begin{cases} 1 & \text{si } t \geq y_k \\ 0 & \text{si } t < y_k. \end{cases}$$

El estimador usual de la función de distribución $F_y(t)$ es el estimador de Horvitz-Thompson dado por:

$$\widehat{F}_{YHT}(t) = \frac{1}{N}\sum_{k\in s} d_k\Delta(t-y_k) \tag{3.2}$$

El estimador $\widehat{F}_{YHT}(t)$ es insesgado bajo el diseño y su varianza viene dada en general por:

$$V(\widehat{F}_{YHT}(t)) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\Delta_{kl}d_k d_l\Delta(t-y_k)\Delta(t-y_l)$$

donde $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$.

En general, el estimador $\widehat{F}_{YHT}(t)$ no es una genuina función de distribución ya que no verifica la propiedad

$$\lim_{t\to+\infty}\widehat{F}_{YHT}(t) = 1$$

Además, el estimador $\widehat{F}_{YHT}(t)$ no incorpora la información auxiliar proporcionada por el vector $\mathbf{x}_k$.

Bajo la condición de que el diseño muestral $p(\cdot)$ sea de tamaño muestral fijo n, la fórmula de Yates-Grundy-Sen es una expresión alternativa de la varianza del estimador $\widehat{F}_{YHT}(t)$, la cual viene dada por:

$$V(\widehat{F}_{YHT}(t)) = -\frac{1}{2}\sum_{k\in s}\sum_{l\in s}\Delta_{kl}(d_k\Delta(t-y_k) - d_l\Delta(t-y_l))^2$$

para la que se dispone del siguiente estimador insesgado:

$$\widehat{V}(\widehat{F}_{YHT}(t)) = -\frac{1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}(d_k\Delta(t-y_k) - d_l\Delta(t-y_l))^2$$

Bajo muestreo aleatorio simple, la varianza del estimador $\widehat{F}_{YHT}(t)$ tiene la siguiente expresión:

$$V(\widehat{F}_{YHT}(t)) = \frac{(1-f)}{N}F_y(t)(1-F_y(t)) \tag{3.3}$$

donde $f = n/N$ es la fracción de muestreo.

La varianza (3.3) ser estimada mediante el siguiente estimador insesgado:

$$\widehat{V}(\widehat{F}_{YHT}(t)) = \frac{(1-f)}{N}\widehat{F}_{YHT}(t)(1-\widehat{F}_{YHT}(t)) \tag{3.4}$$

Adicionalmente, bajo muestreo aleatorio simple el estimador $\widehat{F}_{YHT}(t)$ verifica todas las propiedades de función de distribución.

Para solventar que el estimador $\widehat{F}_{YHT}(t)$ no es una genuina función de distribución, una alternativa para estimar la función de distribución es el estimador de Hájek, definido por:

$$\widehat{F}_{YHJ}(t) = \frac{\sum_{k\in s} d_k \Delta(t-y_k)}{\sum_{k\in s} d_k} \tag{3.5}$$

El estimador $\widehat{F}_{YHJ}(t)$ es consistente y aproximadamente insesgado bajo el diseño (Dorfman, 2009), ya que en general no es insesgado pero su sesgo es reducido y es una aútentica función de distribución al verificar todas las propiedades, pero tampoco incorpora la información auxiliar aportada por el vector $\mathbf{x}_k$.

Bajo muestreo aleatorio simple, el estimador $\widehat{F}_{YHJ}(t)$ coincide con el estimador $\widehat{F}_{YHT}(t)$ y por tanto su varianza viene dada por (3.3) y puede ser estimada mediante 3.4.

3.2. Estimadores indirectos de la función de distribución basados en el diseño

Como se mencionó previamente, existe una gran variedad de estimadores indirectos que permiten la incorporación de la información auxiliar proporcionada por el vector $\mathbf{x}_k$. A continuación revisaremos aquellas técnicas basadas en el diseño que han sido incluidas en los estudios de simulación de los apéndices con propósitos comparativos.

3.2.1. Estimadores de razón y diferencia

Si se asume que la dimensión del vector auxiliar $\mathbf{x}_k$ es $J = 1$, se define el estimador de razón basado en el diseño, de la siguiente manera:

$$\widehat{F}_{YR}(t) = \frac{1}{N}\frac{\sum_{k\in s} d_k \Delta(t-y_k)}{\sum_{k\in s} d_k \Delta(t-\widehat{R}\mathbf{x}_k)} \cdot \sum_{k\in U} d_k \Delta(t-\widehat{R}\mathbf{x}_k) \tag{3.6}$$

donde

$$\widehat{R} = \frac{\sum_{k \in s} d_k y_k}{\sum_{k \in s} d_k \mathbf{x}_k}$$

Si la variable de estudio y es proporcional a la variable auxiliar $\mathbf{x}$ ($y \propto \mathbf{x}$) entonces el estimador $\widehat{F}_{YR}(t)$ proporciona estimaciones perfectas para la la variable de estudio y lo que sugiere que el estimador $\widehat{F}_{YR}(t)$ es más eficiente que los estimadores $\widehat{F}_{YHJ}(t)$ y $\widehat{F}_{YHT}(t)$ cuando y es aproximadamente proporcional a $\mathbf{x}$ (Mukhopadhyay, 2012).

Uno de los inconvenientes del estimador de razón $\widehat{F}_{YR}(t)$ es que no se puede calcular en todos los casos pues es necesario que:

$$\sum_{k \in s} d_k \Delta(t - \widehat{R}\mathbf{x}_k) \neq 0$$

El estimador $\widehat{F}_{YR}(t)$ es asintóticamente insesgado (Rao et al., 1990) y su varianza asintótica viene dada por:

$$V(\widehat{F}_{YR}(t)) = \frac{1}{N^2} V\left(\Delta(t - y_k) - \frac{F_y(t)}{F_x(t/R)} \cdot \Delta(t - R\mathbf{x}_k)\right)$$

donde $R = T_y/T_x$ es el ratio entre los totales poblacionales de las variables $\mathbf{x}$ e y y donde

$$V(y_k) = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s} \Delta_{kl} (d_k y_k - d_l y_l)^2 \tag{3.7}$$

La varianza $V(\widehat{F}_{YR}(t))$ puede ser estimada mediante el siguiente estimador:

$$\widehat{V}(\widehat{F}_{YR}(t)) = \frac{1}{N^2} \widehat{V}\left(\Delta(t - y_k) - \frac{F_y(t)}{F_x(t/\widehat{R})} \cdot \Delta(t - \widehat{R}\mathbf{x}_k)\right)$$

donde

$$\widehat{V}(y_k) = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} (d_k y_k - d_l y_l)^2 \tag{3.8}$$

De manera similar, podemos definir el estimador diferencia de la siguiente manera (Mukhopadhyay, 2012):

$$\widehat{F}_{YD}(t) = \widehat{F}_{YH}(t) + d \cdot \left(\frac{1}{N} \sum_{k \in U} \Delta(t - \widehat{R}\mathbf{x}_k) - \frac{1}{N} \sum_{k \in s} \Delta(t - \widehat{R}\mathbf{x}_k)\right) \tag{3.9}$$

donde d es una constante conocida.

Siguiendo a Rao et al. (1990) el estimador $\widehat{F}_{YD}(t)$ es asintóticamente insesgado y su varianza viene dada

por:

$$V(\widehat{F}_{YD}(t)) = \frac{1}{N^2} V\Big(\Delta(t - y_k) - d \cdot \Delta(t - R\mathbf{x}_k)\Big)$$

donde el operador V viene dado por (3.7).

La varianza del estimador $\widehat{F}_{YD}(t)$ puede ser estimada mediante el siguiente estimador:

$$\widehat{V}(\widehat{F}_{YD}(t)) = \frac{1}{N^2} \widehat{V}\Big(\Delta(t - y_k) - d \cdot \Delta(t - R\mathbf{x}_k)\Big)$$

donde el operador $\widehat{V}$ viene dado por (3.8).

El valor óptimo de d se obtiene minimizando la varianza asintótica del estimador $\widehat{F}_{YD}(t)$. Bajo muestreo aleatorio simple, el valor óptimo de d viene dado por:

$$d_{opt} = \frac{\rho_\Delta S_{\Delta y}}{S_{\Delta x}}$$

con

$$S_{\Delta y} = \frac{1}{N-1} \sum_{k \in U} \Big(\Delta(t - y_k) - F_y(t)\Big)^2$$

$$S_{\Delta x} = \frac{1}{N-1} \sum_{k \in U} \Big(\Delta(t - \widehat{R}\mathbf{x}_k) - F_x(t/\widehat{R})\Big)^2$$

y donde ρ_Δ es el coeficiente de correlación poblacional entre $\Delta(t - y_k)$ y $\Delta(t - \widehat{R}\mathbf{x}_k)$.

En general, la ganancia en eficiencia de los estimadores $\widehat{F}_{YR}(t)$ y $\widehat{F}_{YD}(t)$ sobre los estimadores $\widehat{F}_{YHT}(t)$ y $\widehat{F}_{YHJ}(t)$ suele ser menor que la conseguido con los estimadores de razón y diferencia para la media y totales, ya que la correlación entre $\Delta(t - y_k)$ y $\Delta(t - \widehat{R}\mathbf{x}_k)$ suele ser más débil que la correlación entre y y $\mathbf{x}$.

Si la dimensión J del vector auxiliar $\mathbf{x}_k$ es superior a 1, se pueden definir versiones multivariantes de los estimadores de razón y diferencia (Mukhopadhyay, 2012). Concretamente, el estimador de razón multivariante viene dado por:

$$\widehat{F}_{YR}(t) = \frac{1}{N} \sum_{j=1}^{J} \omega_j \cdot \left(\frac{\sum_{k \in s} d_k \Delta(t - y_k)}{\sum_{k \in s} d_k \Delta(t - \widehat{R}_j \mathbf{x}_{jk})} \right) \cdot \sum_{k \in U} d_k \Delta(t - \widehat{R}_j \mathbf{x}_{jk})$$

donde

$$\widehat{R}_j = \frac{\sum_{k \in s} d_k y_k}{\sum_{k \in s} d_k \mathbf{x}_{jk}}$$

y donde ω_j son constantes a determinar adecuadamente de forma que $\omega_j > 0$ y $\sum_{j=1}^{J} \omega_j = 1$.

De manera similar, el estimador diferencia multivariante se puede definir de la siguiente manera:

$$\widehat{F}_{YD}(t) = \widehat{F}_{YH}(t) + \sum_{j=1}^{J} \omega_j \cdot \left(\frac{1}{N} \sum_{k \in U} \Delta(t - \widehat{R}_j \mathbf{x}_{jk}) - \frac{1}{N} \sum_{k \in s} \Delta(t - \widehat{R}_j \mathbf{x}_{jk}) \right)$$

donde las constantes ω_k, son elegidas optimamente.

Para una revisión más amplia de las versiones multivariantes de los estimadores de razón y diferencia puede consultarse (Mukhopadhyay, 2012).

3.3. Estimadores indirectos de la función de distribución basados en el modelo

De igual forma que con los estimadores basados en el diseño, a continuación revisaremos aquellas técnicas indirectas de estimación de la función de distribución basadas en un modelo con enfoque predictivo que han sido incluidas en los estudios de simulación de los apéndices con propósitos comparativos.

Una vez seleccionada una muestra s, siguiendo a Royall (1970) y Rodrigues et al. (1985), si denotamos por $r = U - s$ las unidades poblacionales no incluidas en la muestra s, los estimadores modelo asistidos o basados en un modelo surgen al considerar que la función de distribución $F_y(t)$ puede descomponerse de la siguiente manera:

$$F_y(t) = \theta_s + \theta_{sr} = \frac{1}{N} \sum_{k \in s} \Delta(t - y_k) + \frac{1}{N} \sum_{k \in r} \Delta(t - y_k) = \theta_s + \theta_r$$

de forma que un estimador para $F_y(t)$ puede obtenerse a partir de un estimador de θ_r, esto es:

$$\widehat{F}_y(t) = \theta_s + \widehat{\theta}_{rs}$$

Para ello, es necesario asumir un modelo de superpoblación entre la variable de estudio y_k y el vector de variables auxiliares $\mathbf{x}_k$, de forma que la eficiencia del estimador obtenido dependerá en gran medida de si el modelo asumido es erróneo o no.

3.3.1. Estimador de Chambers y Dunstan

Chambers & Dunstan (1986) originalmente propusieron un estimador modelo asistido $\widehat{F}_{CD}(t)$ basado en un modelo de regresión donde el vector auxiliar $\mathbf{x}_k$ era de dimensión $J = 1$ y Dorfman (1993) extendió estimador propuesto por Chambers & Dunstan (1986) al caso de un vector auxiliar de dimension $J > 1$. Aquí revisaremos la versión extendida para un modelo de regresión múltiple dado por:

$$y_k = \mathbf{x}_k^{'}\beta + \upsilon_k \epsilon_k \tag{3.10}$$

donde β es un parámetro desconocido, υ_k son constantes conocidas para todas las unidades poblacionales y estrictamente positivas y $\epsilon_k \sim G(0, \sigma^2)$ son variables aleatorias idénticamente distribuidas con función de distribución desconocida G, con media 0 y varianza σ^2.

Para obtener un estimador $\widehat{\theta}_{rs}$, Chambers & Dunstan (1986) estiman $\Delta(t - y_k)$ para todas las unidades $k \in r$. Para ello, tenemos que:

$$E_\xi[\Delta(t - y_k)] = G\left(\frac{t - \mathbf{x}_k^{'}\beta}{\upsilon_k}\right)$$

donde E_ξ denota la esperanza bajo el modelo (3.10). De este modo, podemos estimar $\Delta(t - y_k)$ para $k \in r$ mediante una estimación de $E_\xi[\Delta(t - y_k)]$.

Como consecuencia, se necesita una estimación del parámetro β a partir de la cual podemos estimar los residuos ϵ_j de la siguiente manera:

$$\widehat{\epsilon}_j = \frac{y_j - \mathbf{x}_j^{'}\widehat{\beta}}{\upsilon_j}$$

de forma que un estimador de $\Delta(t - y_k)$ para todas las unidades $k \in r$ viene dado por:

$$\widehat{\Delta}(t - y_k) = \widehat{E}_\xi[\Delta(t - y_k)] = \widehat{G}\left(\frac{t - \mathbf{x}_k^{'}\beta}{\upsilon_k}\right) = \frac{1}{n}\sum_{j \in s}\Delta\left(\frac{t - \mathbf{x}_k^{'}\widehat{\beta}}{\upsilon_k} - \widehat{\epsilon}_j\right) =$$

$$= \frac{1}{n}\sum_{j \in s}\Delta\left(\frac{t - \mathbf{x}_k^{'}\widehat{\beta}}{\upsilon_k} - \frac{y_j - \mathbf{x}_j^{'}\widehat{\beta}}{\upsilon_j}\right)$$

Así, un estimador $\widehat{\theta}_{rs}$ viene dado por:

$$\widehat{\theta}_{rs} = \frac{1}{N \cdot n} \sum_{k \in r} \widehat{\Delta}(t - y_k) = \frac{1}{N} \sum_{k \in r} \sum_{j \in s} \Delta\left(\frac{t - \mathbf{x}_k^{'} \widehat{\beta}}{\upsilon_k} - \frac{y_j - \mathbf{x}_j^{'} \widehat{\beta}}{\upsilon_j} \right)$$

Así, el estimador de *Chambers* y *Dunstan* viene dado por:

$$\widehat{F}_{CD}(t) = \frac{1}{N} \sum_{k \in s} \Delta(t - y_k) + \frac{1}{N} \sum_{k \in r} \sum_{j \in s} \Delta\left(\frac{t - \mathbf{x}_k^{'} \widehat{\beta}}{\upsilon_k} - \frac{y_j - \mathbf{x}_j^{'} \widehat{\beta}}{\upsilon_j} \right) \tag{3.11}$$

Chambers & Dunstan (1986) y Dorfman (1993) consideraron para la estimación de β, el estimador de mínimos cuadrados ponderados $\widehat{\beta}$ dado por:

$$\widehat{\beta} = \left(\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k^{'} / \upsilon_k^2 \right)^{-1} \sum_{k \in s} y_k \mathbf{x}_k / \upsilon_k^2$$

Bajo el modelo de trabajo, el estimador $\widehat{F}_{CD}(t)$ tiene un sesgo insignificante (Dorfman, 2009; Mukhopadhyay, 2012). Para dimensión del vector auxiliar $J = 1$, Chambers et al. (1992) obtuvieron una expresión para la varianza asintótica del estimador $\widehat{F}_{CD}(t)$ asumiendo el siguiente modelo:

$$y_k = a + b \cdot \mathbf{x}_k + \epsilon_k \tag{3.12}$$

donde a y b son parámetros desconocidos y donde ϵ_k son variables aleatorias independientes e idénticamente distribuidas con función de distribución G, media 0 y varianza σ^2 y donde denotaremos por $g = G^{'}$ la función de densidad.

Si asumimos que los valores muestrales y no muestrales de $\mathbf{x}_k$, tiene una función de densidad asintótica común $f(x)$, denotaremos por μ_x y por τ_x^2 la media y varianza de $\mathbf{x}$ respectivamente, esto es:

$$\mu_x = \int x f(x) dx \quad ; \quad \tau_x^2 = \int x^2 f(x) dx - \mu_x^2$$

Para establecer la varianza asintótica del estimador $\widehat{F}_{CD}(t)$ es necesario considerar los siguientes cuatro valores:

$$I_1 = \int (x - \mu_x) g(t - a - b\mathbf{x}) f(x) dx$$

$$I_2 = \int \int G\big((t - a - b\mathbf{x}) \wedge (t - a - b\mathbf{x}^*)\big) f(x) f(x^*) dx dx^*$$

$$I_3 = \int G(t - a - b\mathbf{x}) f(x) dx$$

$$I_4 = \int \left[G(t-a-b\mathbf{x}) - G(t-a-b\mathbf{x})^2\right] f(x)dx$$

donde $a \wedge b = \text{mín}(a, b)$.

Suponiendo que:

$$\lim_{\substack{N\to+\infty \\ n\to+\infty}} f = \lim_{\substack{N\to+\infty \\ n\to+\infty}} \frac{n}{N} = \pi \in (0,1)$$

Con ello, la varianza asintótica de $\widehat{F}_{CD}(t)$ viene dada por:

$$V(\widehat{F}_{CD}(t) - F_y(t)) = \frac{(1-\pi)}{n}\left(\frac{\sigma^2 I_1^2}{\tau_x^2} + I_2 - I_3^2\right) + \frac{1}{N}(1-\pi)I_4 + o(n^{-1}) \tag{3.13}$$

De forma general si el modelo de superpoblación asumido es correcto, el estimador $\widehat{F}_{CD}(t)$ suele ser mucho más eficiente que los estimadores $\widehat{F}_{YHT}(t)$ y $\widehat{F}_{YHT}(t)$ (Dorfman, 2009). No obstante, aun siendo el modelo correcto, el estimador $\widehat{F}_{CD}(t)$ puede experimentar pérdidas en su eficiencia (Chambers et al., 1992). Por otro lado, si el modelo de regresión adoptado no es adecuado, el estimador $\widehat{F}_{CD}(t)$ sufre un aumento de sesgo considerable.

3.3.2. Estimador de Rao–Kovar–Mantel

Para solventar los inconvenientes del estimador $\widehat{F}_{CD}(t)$ cuando el modelo es incorrecto, Rao et al. (1990) propusieron un estimador de tipo diferencia $\widehat{F}_{RKM}(t)$ que es consistente bajo el diseño. Para ello, bajo el modelo (3.10)se define la siguiente variable auxiliar:

$$G_k = \frac{1}{N}\sum_{j\in U} \Delta\left(\frac{(t - \mathbf{x}_k^{'}\beta)}{\upsilon_k} - V_{nj}\right)$$

donde V_{nj} se define de la siguiente manera:

$$V_{nj} = \frac{(y_j - \mathbf{x}_j^{'}\beta)}{\upsilon_j}$$

A partir de la variable auxiliar G_k, podemos considerar el siguiente estimador diferencia:

$$\tilde{F}_{RKM}(t) = \widehat{F}_{YHT}(t) + \frac{1}{N}\left(\sum_{k\in U} G_k - \sum_{k\in s} d_k G_k\right)$$

El estimador $\tilde{F}_{RKM}(t)$ es insesgado bajo el diseño y también es asintóticamente insesgado bajo el modelo (Mukhopadhyay, 2012), pero dado que los valores G_k son desconocidos en general, es necesario estimarlos. Para ello, vamos a considerar el estimador de mínimos cuadrados ponderados $\widehat{\beta}_\pi$ mediante los pesos d_k

(Dorfman, 2009).

En primer lugar, para la estimación de los valores G_k para el término $\sum_{k\in U} G_k$, vamos a considerar el siguiente estimador:

$$\widehat{G}_{\pi k} = \frac{\sum_{j\in s} d_j \Delta\Big(\frac{(t - \mathbf{x}_k^{'}\widehat{\beta}_\pi)}{\upsilon_k} - \widehat{V}_{nj}\Big)}{\sum_{j\in s} d_j}$$

donde

$$\widehat{V}_{nj} = \frac{(y_j - \mathbf{x}_j^{'}\widehat{\beta}_\pi)}{\upsilon_j}$$

De forma similar, para la estimación de G_k en el término $\sum_{k\in s} d_k G_k$, se considera el siguiente estimador:

$$\widehat{G}_{\pi ck} = \frac{1}{\sum_{j\in s} \pi_k/\pi_{kj}} \sum_{j\in s} \frac{\pi_k}{\pi_{kj}} \Delta\Big(\frac{(t - \mathbf{x}_k^{'}\widehat{\beta}_\pi)}{\upsilon_k} - \widehat{V}_{nj}\Big)$$

Finalmente, el estimador de Rao–Kovar–Mantel viene dado por:

$$\widehat{F}_{RKM}(t) = \widehat{F}_{YHT}(t) + \frac{1}{N}\Big(\sum_{k\in U} \widehat{G}_{\pi k} - \sum_{k\in s} d_k \widehat{G}_{\pi ck}\Big) \tag{3.14}$$

el cual es tanto asintóticamente insesgado bajo el diseño como bajo el modelo (Mukhopadhyay, 2012).

Al incorporar probabilidades de inclusión de segundo orden, el estimador $\widehat{F}_{RKM}(t)$ es complejo si consideramos diseños muestrales diferentes al muestreo aleatorio simple o estratificado (Dorfman, 2009) y en general tiene un alto coste computacional (Mukhopadhyay, 2012).

Bajo el modelo (3.12), Chambers et al. (1992) establecieron la varianza asintótica del estimador $\widehat{F}_{RKM}(t)$ que viene dada por:

$$V(\widehat{F}_{RKM}(t) - F_y(t)) = \frac{1}{n}(1-\pi)I_4 + \frac{1}{N}(1-\pi)I_4 + o(n^{-1}) \tag{3.15}$$

donde $\pi \in (0, 1)$ y I_4 fueron definidas previamente.

En general, aunque el modelo sea aproximadamente correcto, el estimador $\widehat{F}_{RKM}(t)$ tenderá a ofrecer un mejor comportamiento que $\widehat{F}_{YHT}(t)$ y $\widehat{F}_{YHJ}(t)$ (Dorfman, 2009). Adicionalmente, si el modelo considerado es erróneo, el estimador $\widehat{F}_{RKM}(t)$ tiende a comportarse mejor que $\widehat{F}_{CD}(t)$. Sin embargo, si el modelo asumido es correcto entonces el estimador $\widehat{F}_{CD}(t)$ puede ser notablemente más eficiente que $\widehat{F}_{RKM}(t)$

(Dorfman, 2009). Dado que $I_2 - I_3^2 \leq I_4$ (Chambers et al., 1992), si no tenemos en cuenta el término asociado a I_1, tendríamos que:

$$V(\widehat{F}_{CD}(t) - F_y(t)) \leq V(\widehat{F}_{RKM}(t) - F_y(t))$$

Ahora bien, aunque el modelo asumido sea correcto, debido al término asociado a I_1, incluso cuando el modelo es correcto, $\widehat{F}_{CD}(t)$ puede ofrecer un mal comportamiento para algunos valores de t (Chambers et al., 1992).
De este modo, el diagnóstico del modelo es esencial a la hora de la estimación de la función de distribución a través de los estimadores $\widehat{F}_{CD}(t)$ y $\widehat{F}_{RKM}(t)$. Así, el estimador $\widehat{F}_{RKM}(t)$ se adapta mejor a una incorrecta especificación del modelo, de forma que ofrece mejores resultados que $\widehat{F}_{CD}(t)$ mientras que si el modelo es correcto, el estimador $\widehat{F}_{CD}(t)$ muestra en egeneral un mejor comportamiento que $\widehat{F}_{RKM}(t)$ (Dorfman, 2009).

Para una mayor revisión de los estimadores $\widehat{F}_{CD}(t)$ y $\widehat{F}_{RKM}(t)$ puede consultarse Chambers & Dunstan (1986); Chambers et al. (1992); Mukhopadhyay (2012); Dorfman (1993, 2009) y Rao et al. (1990).

3.4. Estimadores calibrados para la función de distribución

A continuación revisaremos los estimadores basados en el método de calibración para la función de distribución necesarios para alcanzar los objetivos planteados en la presente tesis doctoral. Primeramente, realizaremos una revisión del método de calibración, que originalmente fue desarrollado por Deville & Särndal (1992) para la estimación de totales y medias. Seguidamente, revisaremos el estimador calibrado para $F_y(t)$ desarrollado por Rueda et al. (2007a) y sus propiedades más relevantes. Finalmente, dado que el comportamiento asintótico del estimador de Rueda et al. (2007a) depende de la elección de un vector de puntos auxiliares, también revisaremos las versiones óptimas de este estimador bajo muestreo aleatorio simple desarrolladas en Martínez et al. (2010, 2015) y Martínez et al. (2017).

3.4.1. Método de calibración

El método de calibración fue desarrollado originalmente por Deville & Särndal (1992) para la estimación de totales y medias como se ha mencionado anteriormente y ha sido un método bastante considerado en la literatura previa sobre investigaciones muestrales a la hora de desarrollar nuevos estimadores para el total poblacional T_y de una variable de estudio y o la media $\bar{Y}$ de dicha variable (Arcos et al., 2014; Brewer, 2000; Estevao & Särndal, 2006, 2000; Kim & Park, 2010; Montanari & Ranalli, 2005; Ranalli et al., 2016; Rueda

et al., 2006, 2009; Wu & Sitter, 2001).

A continuación, pasamos a revisar el método desarrollado por Deville & Särndal (1992). Supongamos que estamos interesados en estimar el total T_y de la variable y, dado por:

$$T_y = \sum_{k \in U} y_k$$

Si una muestra s es obtenida a partir de un diseño muestral $p(\cdot)$, con pesos básicos dados por d_k, el estimador usual de T_y es el estimador de Horvitz-Thompson dado por:

$$T_{YH} = \sum_{k \in s} d_k y_k \tag{3.16}$$

El estimador T_{YH} es insesgado para T_y pero no incorpora la información auxiliar disponible a través del vector auxiliar $\mathbf{x}_k$. Para incorporar esta información Deville & Särndal (1992) propusieron sustituir los pesos básicos d_k por unos nuevos pesos ω_k que esten lo más próximo posible a los pesos básicos d_k respecto a una medida de distancia de forma que proporcionen estimaciones perfectas para el total $T_{\mathbf{x}}$ del vector auxiliar $\mathbf{x}_k$, esto es:

$$\sum_{k \in s} \omega_k \mathbf{x}_k = T_{\mathbf{x}} \tag{3.17}$$

Para ello, Deville & Särndal (1992) consideraron la minimización de distancia chi-cuadrado entre los pesos básicos d_k y los nuevos pesos calibrados ω_k, dada por:

$$\Phi_s = \frac{1}{2} \sum_{k \in s} \frac{(\omega_k - d_k)^2}{d_k q_k} \tag{3.18}$$

sujeta a la condición (3.17) y donde q_k son constantes positivas y conocidas.

Para ello, aplicando el método de multiplicadores de Lagrange, debemos minimizar:

$$\Theta_s(\omega_k) = \sum_{k \in s} \frac{(\omega_k - d_k)^2}{d_k q_k} + \lambda' \cdot \left(T_{\mathbf{x}} - \sum_{k \in s} \omega_k \mathbf{x}_k \right)$$

donde λ es el vector de multiplicadores de Lagrange.

Derivando parcialmente $\Theta_s(\omega_k)$ con respecto a ω_k, se obtiene el siguiente sistema de ecuaciones:

$$\frac{\partial \Theta_s(\omega_k)}{\partial \omega_k} = \frac{(\omega_k - d_k)}{d_k q_k} - \lambda^{'} \cdot \mathbf{x}_k = 0$$

lo que lleva a los pesos calibrados:

$$\omega_k = d_k + d_k q_k + \lambda^{'} \cdot \mathbf{x}_k$$

Dado que los pesos calibrados ω_k deben satisfacer la condición (3.17), tenemos que:

$$T_{\mathbf{x}}^{'} = \sum_{k \in s} (d_k + d_k q_k + \lambda^{'} \cdot \mathbf{x}_k)\mathbf{x}_k^{'}$$

con lo que el vector λ viene dado por:

$$\lambda = T_s^{-1}(T_{\mathbf{x}} - \widehat{T}_{XH})$$

donde $\widehat{T}_{XH}$ denota el estimador de Horvitz-Thompson para estimar el total $T_{\mathbf{x}}$ y supuesto que la matriz T_s dada por

$$T_s = \sum_{k \in s} d_k q_k \mathbf{x}_k \mathbf{x}_k^{'}$$

es no singular.

Los pesos calibrados resultantes del proceso de calibración vienen dados por:

$$\omega_k = d_k + d_k q_k + (T_{\mathbf{x}} - \widehat{T}_{XH})^{'} \cdot T_s^{-1} \cdot \mathbf{x}_k \tag{3.19}$$

Finalmente, el estimador calibrado $\widehat{T}_{yc}$ para el total T_y obtenido con los pesos calibrados (3.19) es:

$$\widehat{T}_{yc} = \widehat{T}_{YH} + (T_{\mathbf{x}} - \widehat{T}_{XH})^{'} \cdot \widehat{B}_s \tag{3.20}$$

donde

$$\widehat{B}_s = T_s^{-1} \cdot \sum_{k \in s} d_k q_k \mathbf{x}_k y_k$$

de forma que el estimador calibrado resultante $\widehat{F}_{yc}(t)$ es el estimador general de regresión (Särndal, 1980).

Deville & Särndal (1992) demostraron que el estimador calibrado $\widehat{T}_{yc}$ es asintóticamente insesgado y su varianza asintótica viene dada por:

$$AV(\widehat{T}_{yc}) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl}(d_k U_k)(d_l U_l) \tag{3.21}$$

donde $U_k = y_k - \mathbf{x}_k^{'} \cdot B$ con

$$B = \left(\sum_{k \in U} q_k \mathbf{x}_k \mathbf{x}_k^{'} \right)^{-1} \cdot \left(\sum_{k \in U} q_k \mathbf{x}_k y_k \right) \tag{3.22}$$

La varianza (3.21) puede ser estimada mediante el siguiente estimador (Deville & Särndal, 1992):

$$\widehat{V}(\widehat{T}_{yc}) = \sum_{k \in U} \sum_{l \in U} \frac{\Delta_{kl}}{\pi_{kl}} (d_k u_k)(d_l u_l)$$

con $u_k = y_k - \mathbf{x}_k^{'} \cdot \widehat{B}_s$.

Un estimador alternativo de la varianza del estimador $\widehat{T}_{yc}$ puede obtenerse al reemplazar los pesos básicos d_k por los pesos calibrados ω_k dados por (3.19) en el estimador $\widehat{V}(\widehat{T}_{yc})$.

Deville & Särndal (1992) consideraron una familia de distancias para el desarrollo del proceso de calibración descrito. Concretamente, Deville & Särndal (1992) consideraron para cada unidad k, una medida de distancia $G_k(\omega, d)$ de forma que para cada valor fijo $d > 0$:

- $G_k(\omega, d)$ es una función no negativa.
- $G_k(\omega, d)$ es diferenciable con respecto a ω.
- $G_k(\omega, d)$ es estrictamente convexa.
- $G_k(\omega, d)$ está definida en un intervalo $D_k(d)$ que contiene a d de forma que $G_k(d, d) = 0$
- La derivada parcial $g_k(\omega, d) = \dfrac{\partial G_k(\omega, d)}{\partial \omega}$ es una función continua y biyectiva en el intervalo $D_k(d)$ y con espacio imagen $Im_k(d)$.

La minimización de la distancia

$$\sum_{k \in s} G_k(\omega_k, d_k)$$

sujeto a la condición de calibración (3.17) lleva a los pesos calibrados:

$$\omega_k = d_k F_k(\mathbf{x}_k^{'} \cdot \lambda)$$

donde $d_k F_k(\cdot)$ es la función inversa de la función $g_k(\cdot, d_k)$.

Ejemplos de distancias pertenencientes a esta familia de distancias pueden consultarse en Deville & Särndal (1992) siendo la distancia chi-cuadrado dada por (3.18) una de las distancias incluidas en esta

familia. Deville & Särndal (1992) demostraron que los estimadores calibrados obtenidos con una distancia pertenenciente a esta familia de distancias tienen el mismo comportamiento asintótico que el estimador $\widehat{T}_{yc}$ y por tanto comparte la varianza asintótica dada por (3.21).

Para una revisión más amplia del método de calibración pueden consultarse Deville & Särndal (1992) y Kim & Park (2010).

3.4.2. Estimador calibrado de Rueda et al. (2007a) para $\mathbf{F_y(t)}$

El método de calibración anteriormente descrito, ha sido adaptado desde diferentes perspectivas a la estimación de la función de distribución en diversos estudios previos obteniéndose nuevos estimadores (Breidt et al., 2007; Harms & Duchesne, 2006; Kovacevic, 1997; Mayor-Gallego et al., 2019; Rueda et al., 2007a; Wu, 2003).
La propuesta de Rueda et al. (2007a) destaca por ser computacionalmente sencilla (Särndal, 2007) y se basa en el método de calibración para minimizar la distancia chi-cuadrado sujeta a restricciones que requieren la consideración de P puntos de la variable auxiliar elegidos arbitrariamente.

Concretamente, la propuesta de Rueda et al. (2007a) se basa en la definición de la siguiente pseudo-variable auxiliar g_k a partir del vector auxiliar $\mathbf{x}_k$:

$$g_k = \widehat{\beta}' \mathbf{x}_k \text{ for } k = 1, 2, ...N \tag{3.23}$$

$$\widehat{\beta} = \left(\sum_{k \in s} d_k \mathbf{x}_k \mathbf{x}_k^{'} \right)^{-1} \cdot \sum_{k \in s} d_k \mathbf{x}_k y_k \tag{3.24}$$

A partir de la pseudo-variable g_k, la propuesta de Rueda et al. (2007a) consiste en reemplazar los pesos básicos d_k empleados en el estimador de Horvitz-Thompson $F_{YHT}(t)$ por unos nuevos pesos calibrados ω_k de forma que minimicen la distanci chi-cuadrado dada por (4.5 sujeta las siguientes restricciones de calibración:

$$\frac{1}{N} \sum_{k \in s} \omega_k \Delta(t_j - g_k) = F_g(t_j) \quad j = 1, 2, \ldots, P \tag{3.25}$$

donde $t_1, t_2, \ldots, t_P$ son puntos arbitrariamente elegidos, de forma que supondremos sin pérdida de generabilidad, que:

$$t_1 < t_2 < \ldots t_P$$

y donde $F_g(t_j)$ denota la función de distribución de la pseudo-variable g_k evaluada en los puntos t_j, $\quad j =$

$1, 2, \ldots, P$.

De este modo, con los nuevos pesos calibrados ω_k, lo que se prentende es obtener un estimador calibrado

$$\widehat{F}_{yc}(t) = \frac{1}{N}\sum_{k\in s}\omega_k\Delta(t - y_k)$$

de forma que:

- $\widehat{F}_{yc}(t)$ es asintóticamente insesgado.
- $\widehat{F}_{yc}(t)$ proporcione estimación exactas para la función de distribución $F_g(t_j)$ de la pseudo-variable g evaluada en los puntos $t_j, \quad j = 1, 2, \ldots, P$.

Si denotamos por:

$$\mathbf{t_g}^{'} = (t_1, t_2, \ldots, t_P)$$

$$\Delta(\mathbf{t_g} - g_k)^{'} = (\Delta(t_1 - g_k), \Delta(t_2 - g_k), \ldots, \Delta(t_P - g_k))$$

$$F_g(\mathbf{t_g})^{'} = \left(F_g(t_1), F_g(t_2), \ldots, F_g(t_P)\right)$$

$$\widehat{F}_{GHT}(\mathbf{t_g})^{'} = \left(\widehat{F}_{GHT}(t_1), \widehat{F}_{GHT}(t_2), \ldots, \widehat{F}_{GHT}(t_P)\right)$$

y a través de un proceso similar al desarrollado con el método de calibración para la estimación del total T_y anteriormente descrito donde en este caso es asumido que la matriz T_s dada por:

$$T_s = \sum_{k\in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)'$$

es no singular, los pesos calibrados obtenidos son:

$$\omega_k = d_k + d_k q_k \left(F_g(\mathbf{t_g}) - \widehat{F}_{GHT}(\mathbf{t_g})\right)^{'} \cdot T_s^{-1} \cdot \Delta(\mathbf{t_g} - g_k) \tag{3.26}$$

donde $\widehat{F}_{GHT}(\mathbf{t_g})$ denota el estimador de Horvitz-Thompson estimator para la función de distribución de la pseudo-variable $F_g(\mathbf{t_g})$ evaluada en $\mathbf{t_g} = (t_1, \ldots, t_P)^{'}$.

A partir de los pesos (3.26), el estimador calibrado de Rueda et al. (2007a) viene dado por:

$$\widehat{F}_{yc}(t) = \frac{1}{N}\sum_{k\in s}\omega_k\Delta(t - y_k) = \widehat{F}_{YHT}(t) + \left(F_g(\mathbf{t_g}) - \widehat{F}_{GHT}(\mathbf{t_g})\right)^{'} \cdot \widehat{D}(\mathbf{t_g}) \tag{3.27}$$

donde

$$\widehat{D}(\mathbf{t_g}) = T^{-1} \cdot \sum_{k \in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(t - y_k)$$

Rueda et al. (2007a) analizaron las condiciones a partir de las cuales se puede garantizar la existencia de la matriz T_s^{-1}. Para ello, debemos considerar los valores muestrales de la pseudo-variable g ordenados ascendentemente

$$g_{(1)} \leq g_{(2)} \leq \cdots g_{(n)}$$

y denotaremos por k_j el número de unidades muestrales cuyo valor de la pseudo-variable g es inferior o igual a t_j.

Rueda et al. (2007a) demostraron que la existencia de T_s^{-1} está garantizada si asumimos que $k_j < k_{j+1}$ para $j = 1, 2, \ldots, P-1$ y $k_1 > 0$, en cuyo caso la matriz T_s^{-1} es una matriz simétrica que tiene la forma:

$$T_s^{-1} = \begin{pmatrix} a_{11} & a_{12} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \cdots & 0 \\ 0 & a_{32} & a_{33} & a_{34} & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & a_{PP-1} & a_{P-1P} \\ 0 & 0 & 0 & 0 & a_{P-1P-1} & a_{P-1P} \end{pmatrix}$$

donde los valores a_{ij} para $i = 1, 2, \ldots, P-1$ se definen de la siguiente manera:

$$a_{ii} = \frac{1}{\sum_{k=k_{i-1}+1}^{k_i} d_k q_k} + \frac{1}{\sum_{k=k_i+1}^{k_{i+1}} d_k q_k}$$

$$a_{ii+1} = -\frac{1}{\sum_{k=k_i+1}^{k_{i+1}} d_k q_k}$$

con $a_{ij} = 0$ si $j \neq i-1$, $j \neq i$ y $j \neq i+1$, y k_0 se establece igual a 1. Finalmente, para $i = P$, tenemos que:

$$a_{PP} = \frac{1}{\sum_{k=k_{P-1}+1}^{k_P} d_k q_k}$$

Con esta nueva expresión de la matriz T_s^{-1}, el estimador calibrado de Rueda et al. (2007a) puede expresarse

de la siguiente manera:

$$\widehat{F}_{yc}(t) = \widehat{F}_{YHT}(t) + \sum_{i=1}^{P} \left(F_g(t_i) - \widehat{F}_{GHT}(t_i) \right)' \cdot A_i$$

con

$$A_i = \frac{\sum_{k=k_{i-1}+1}^{k_i} d_k q_k \Delta(t - y_{(k)})}{\sum_{k=k_{i-1}+1}^{k_i} d_k q_k} - \frac{\sum_{k=k_i+1}^{k_{i+1}} d_k q_k \Delta(t - y_{(k)})}{\sum_{k=k_i+1}^{k_{i+1}} d_k q_k} \quad i = 1, 2, \ldots, P-1$$

$$A_P = \frac{\sum_{k=k_{P-1}+1}^{k_P} d_k q_k \Delta(t - y_{(k)})}{\sum_{k=k_{i-1}+1}^{k_i} d_k q_k}$$

El estimador $\widehat{F}_{yc}(t)$ es asintóticamente insesgado y su varianza asintótica viene dada por:

$$V(\widehat{F}_{yc}(t)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (d_k E_k)(d_l E_l) \tag{3.28}$$

con $E_k = \Delta(t - y_k) - \Delta(\mathbf{t_g} - g_k) \cdot D(\mathbf{t_g})$ y donde

$$D(\mathbf{t_g}) = \left(\sum_{k \in U} q_k \Delta(\mathbf{t_g} - g_k) \Delta(\mathbf{t_g} - g_k)' \right)^{-1} \cdot \left(\sum_{k \in U} q_k \Delta(\mathbf{t_g} - g_k) \Delta(t - y_k) \right). \tag{3.29}$$

Como consecuencia, el comportamiento asintótico $\widehat{F}_{yc}(t)$ depende de la elección del vector $\mathbf{t_g}$ y por tanto su precisión se ve influenciada por dicha elección. Un estimador para la varianza asintótica (3.28) del estimador $\widehat{F}_{yc}(t)$ viene dado por:

$$\widehat{V}(\widehat{F}_{yc}(t)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \frac{\Delta_{kl}}{\pi_{kl}} (d_k e_k)(d_l e_l) \tag{3.30}$$

donde $e_k = \Delta(t - y_k) - \Delta(\mathbf{t_g} - g_k) \cdot \widehat{D}(\mathbf{t_g})$.

Como veremos más adelante, bajo condiciones suaves el estimador $\widehat{F}_{yc}(t)$ es una auténtica función de distribución con lo que podríamos emplearlo directamente en la estimación de cuantiles y medidas de pobreza. Ahora bien, como se comentó anteriormente, en los estudios de pobreza es habitual que se produzca falta de respuesta por lo que es necesario adaptar la metodología propuesta en Rueda et al. (2007a) para el tratamiento de falta de respuesta. En el Apéndice 1, se desarrollan técnicas de calibración para estimar la función de distribución bajo falta de respuesta mientras que en el Apéndice 2 estas técnicas serán adaptadas

a la estimación de cuantiles y medidas de pobreza.

3.4.3. Estimador calibrado óptimo de Martínez et al. (2017) para $\mathbf{F_y(t)}$

Dado que el comportamiento asintótico del estimador $\widehat{F}_{yc}(t)$ depende de la elección de puntos auxiliares $\mathbf{t}_g$, es necesario buscar la elección óptima de dicho vector para mejorar la eficiencia del estimador $\widehat{F}_{yc}(t)$, esto es, la elección del vector $\mathbf{t}_g$ que minimiza la varianza asintótica del estimador $\widehat{F}_{yc}(t)$ dada por (3.28). Así, bajo muestreo aleatorio simple, Martínez et al. (2010) determinaron la elección óptima del vector $\mathbf{t}_g$ en el caso de que la dimensión del mismo sea $P = 1$. De igual forma, bajo muestreo aleatorio simple Martínez et al. (2012) extendieron los resultados de Martínez et al. (2010) para el caso de dimensión $P = 2$ y posteriormente Martínez et al. (2015) trataron el caso general donde la dimensión P puede ser mayor que 2. A pesar de establecerse en Martínez et al. (2015) la elección óptima del estimador para una dimensión P cualquiera, todavía quedaba por resolver cuál es la dimensión óptima del vector $\mathbf{t}_g$ para minimizar la varianza asintótica del estimador $\widehat{F}_{yc}(t)$. Finalmente, Martínez et al. (2017), también bajo muestreo aleatior simple, obtuvieron la dimensión óptima del vector $\mathbf{t}_g$, así como la elección óptima de $\mathbf{t}_g$.

Para la consecución de los objetivos de la presente tesis, revisaremos el estimador calibrado de Martínez et al. (2017) basado en la elección óptima de la dimensión del vector $\mathbf{t}_g$, para lo que también será necesario revisar la elección óptima abordada por Martínez et al. (2015).

Siguiendo a Martínez et al. (2015), el estimador $\widehat{F}_{yc}(t)$ puede expresarse de la siguiente manera:

$$\widehat{F}_{yc}(t) = \widehat{F}_{YHT}(t) + \sum_{i=1}^{P} \left(F_g(t_i) - \widehat{F}_{GHT}(t_i)\right) D_i$$

y en consecuencia la varianza asintótica (3.28) del estimador $\widehat{F}_{yc}(t)$ puede expresarse de la siguiente manera:

$$V(\widehat{F}_{yc}(t)) = V(\widehat{F}_{YHT}(t)) + \sum_{i=1}^{P} D_j^2 V(\widehat{F}_{GHT}(t_i)) +$$

$$+2 \sum_{j<i} D_j D_i Cov(\widehat{F}_{GHT}(t_j), \widehat{F}_{GHT}(t_i)) - 2 \sum_{i=1}^{P} D_i Cov(\widehat{F}_{YHT}(t), \widehat{F}_{GHT}(t_i))$$

donde el vector de coeficientes $D(\mathbf{t_g}) = (D_1, D_2, \ldots, D_P)$ viene dado por (4.9).
En consecuencia, la minimización de la varianza $V(\widehat{F}_{yc}(t))$ respecto al vector $\mathbf{t_g}' = (t_1, t_2, \ldots, t_P)$ equivale

a la minimización de la siguiente función:

$$\begin{aligned} G(t_1, t_2, \ldots, t_P) &= \sum_{i=1}^{P} D_j^2 V(\widehat{F}_{GHT}(t_i)) + 2 \sum_{j<i} D_j D_i Cov(\widehat{F}_{GHT}(t_j), \widehat{F}_{GHT}(t_i)) \\ &- 2 \sum_{i=1}^{P} D_i Cov(\widehat{F}_{YHT}(t), \widehat{F}_{GHT}(t_i)) \end{aligned} \tag{3.31}$$

Si consideramos $q_k = 1$ para todo $k \in U$ y teniendo en cuenta que $t_1 < t_2 \ldots < t_P$, la matriz simétrica $\sum_{k \in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)'$ viene dada por:

$$\sum_{k \in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)' = \sum_{k \in U} q_k \Delta(t_i - g_k)\Delta(t_j - g_k) = A_{ij} \tag{3.32}$$

con

$$A_{ij} = N F_g(t_i) \quad \text{si } i \leq j$$

Puesto que el vector de coeficientes $D(\mathbf{t_g})$ depende de la matriz inversa de (3.32), se puede demostrar siguiendo a Martínez et al. (2015) que la inversa de (3.32) es una matriz simétrica $C = C_{ij}$ dada por:

$$C_{ii} = \frac{1}{N}\left[\frac{1}{\left(F_g(t_i) - F_g(t_{i-1})\right)} + \frac{1}{\left(F_g(t_{i+1}) - F_g(t_i)\right)}\right]$$

$$C_{ii+1} = \frac{-1}{N}\left[\frac{1}{\left(F_g(t_{i+1}) - F_g(t_i)\right)}\right]$$

para $i = 1, 2, \ldots, P-1$ de forma que $C_{ij} = 0$ para $j > i+1$ y donde es necesario establecer que $F_g(t_0) = 0$. Finalmente, para $i = P$, tenemos que:

$$C_{PP} = \frac{1}{N}\frac{1}{\left(F_g(t_P) - F_g(t_{P-1})\right)}$$

Con ello, si denotamos por:

$$K_i = \sum_{k \in U} \Delta(t - y_k)\Delta(t_i - g_k) \quad i = 1, 2, \ldots, P$$

y definimos $K_0 = 0$, el vector de coeficientes $D(\mathbf{t_g})$ viene dado por:

$$D_i = \frac{1}{N}\left[\frac{(K_i - K_{i-1})}{\left(F_g(t_i) - F_g(t_{i-1})\right)} - \frac{(K_{i+1} - K_i)}{\left(F_g(t_{i+1}) - F_g(t_i)\right)}\right] \quad i = 1, 2, \ldots, P-1 \tag{3.33}$$

$$D_P = \frac{1}{N}\frac{(K_P - K_{P-1})}{\big(F_g(t_P) - F_g(t_{P-1})\big)} \tag{3.34}$$

Dado que bajo muestreo aleatorio simple, tenemos que:

$$V(\widehat{F}_{YHT}(t)) = \frac{N}{N-1}F_y(t)\big(1 - F_y(t)\big) \tag{3.35}$$

$$V(\widehat{F}_{GHT}(t_i)) = \frac{N}{N-1}F_g(t_i)\big(1 - F_g(t_i)\big) \quad i = 1, 2, \ldots, P \tag{3.36}$$

$$Cov(\widehat{F}_{YHT}(t), \widehat{F}_{GHT}(t_i)) = \frac{1}{N-1}\big[K_i - NF_y(t)F_g(t_i)\big] \quad i = 1, 2, \ldots, P \tag{3.37}$$

$$Cov(\widehat{F}_{GHT}(t_j), \widehat{F}_{GHT}(t_i)) = \frac{N}{N-1}F_g(t_i)\big(1 - F_g(t_j)\big) \quad i > j \tag{3.38}$$

con lo que la función $G(t_1, t_2, \ldots, t_P)$ puede expresarse como:

$$G(t_1, t_2, \ldots, t_P) = 2NF_y(t)K_P - \sum_{i=1}^{P}\frac{(K_i - K_{i-1})^2}{\big(F_g(t_i) - F_g(t_{i-1})\big)} - K_P^2 \tag{3.39}$$

Si consideramos los conjuntos A_t y B_t dados respectivamente por:

$$A_t = \{g_k : k \in U; y_k \leq t\} = \{a_1^t, a_2^t, \ldots, a_{M_t}^t\} \tag{3.40}$$

con $a_1^t < a_2^t \ldots < a_{M_t}^t$ y donde

$$B_t = \{b_1^t, b_2^t, \ldots, b_{M_t}^t\} \tag{3.41}$$

de forma que:

$$b_1^t = \max_{k \in U_1}\{g_k\} \quad \text{con } U_1 = \{k \in U : g_k < a_1^t\}$$

$$b_l^t = \max_{k \in U_l}\{g_k\} \quad \text{con } U_l = \{k \in U : a_{l-1}^t < g_k < a_1^t\} \quad l = 2, \ldots, M_t$$

de forma que si suponemos que b_l^t existe para todo $l = 1, 2, \ldots, M_t$, tenemos que:

$$b_1^t < a_1^t < b_2^t < a_2^t < \ldots < b_{M_t}^t < a_{M_t}^t$$

Según Martínez et al. (2015), el mínimo global de la función (3.39) se alcanza en un vector $\mathbf{t_{opt}} = (t_1, t_2, \ldots, t_P)$ donde $t_i \in A_t$ o bien $t_i \in B_t$. En el caso de que algún valor b_l^t no exista entonces el problema de minimización es más simple que el caso donde todos los valores b_l^t existen.

Dado que los conjuntos A_t y B_t no son conocidos pues dependen de los valores poblacionales de la variable de estudio y que sólo es conocida para las unidades muestrales, el vector óptimo $\mathbf{t_{opt}}$ no se puede

obtener y en consecuencia tampoco el estimador calibrado basado en dicha elección óptima. Incluso si el vector óptimo $\mathbf{t_{opt}}$ fuese conocido, en algunos casos no sería posible obtener el estimador calibrado, ya que podrían surgir restricciones de calibración incompatibles al emplearlo en la ecuación (4.6) (Martínez et al., 2017). Por ello, es necesario recurrir a versiones muestrales de los conjuntos A_t y B_t que permitan obtener una estimación del vector óptimo $\mathbf{t_{opt}}$. Así, si consideramos:

$$A_{st} = \{g_k : k \in s; y_k \leq t\} = \{a^t_{1s}, a^t_{2s}, \ldots, a^t_{ms_t}\} \tag{3.42}$$

con $a^t_{1s} < a^t_{2s} \ldots < a^t_{ms_t}$ y

$$B_{st} = \{b^t_{1s}, b^t_{2s}, \ldots, b^t_{ms_t}\} \tag{3.43}$$

donde

$$b^t_{1s} = \underset{k \in U_{1s}}{\text{máx}} \{g_k\} \quad \text{con } U_{1s} = \{k \in s \,:\, g_k < a^t_{1s}\}$$

$$b^t_{ls} = \underset{k \in U_{ls}}{\text{máx}} \{g_k\} \quad \text{con } U_{ls} = \{k \in s \,:\, a^t_{(l-1)s} < g_k < a^t_{1s}\} \quad l = 2, \ldots, ms^t_t$$

Así, basandonos en una versión muestral de la función $G(t_1, t_2, \ldots, t_P)$, que vendría dada por:

$$G_s(t_1, t_2, \ldots, t_P) = 2N\widehat{F}_{YH}(t)K_{Ps} - \sum_{i=1}^{P} \frac{(K_{is} - K_{(i-1)s})^2}{(\widehat{F}_{GHT}(t_i) - \widehat{F}_{GHT}(t_{i-1})} - K^2_{Ps} \tag{3.44}$$

podemos determinar una estimación del vector óptimo $\mathbf{t^s_{opt}} = (t_{1s}, t_{2s}, \ldots, t_{PS})$ de forma que $t_{is} \in A_{st}$ o bien $t_{is} \in B_{st}$.

Ahora bien, aunque Martínez et al. (2015) determinaron el conjunto de posibles candidatos del vector óptimo para una dimensión fija P, no determinaron cuál debe ser el valor de P óptimo ni el vector óptimo asociado al valor óptimo de P. Martínez et al. (2017) sí determinaron la dimensión óptima P del vector $\mathbf{t_g}$, así como su elección óptima sin necesidad de seleccionarlo de un posible conjunto de candidatos. Concretamente, si consideramos los conjuntos A_t y B_t dados respectivamente por (3.40) y (3.41), tenemos que el vector óptimo auxiliar $\mathbf{t}_g$ tiene dimensión óptima $P = 2M_t$ si b^t_l existe para todo $l = 1, \ldots, M_t$, en cuyo caso el vector óptimo $\mathbf{t}_{OPT}$ viene dado por:

$$\mathbf{t_{OPT}}(t) = (b^t_1, a^t_1, \ldots, b^t_{M_t}, a^t_{M_t}). \tag{3.45}$$

Si para agunos valores $l^t_1, l^t_2, \ldots l^t_{p_t} \in \{1, \ldots, M_t\}$; no existe $b^t_{l_h}$ con $h = 1, 2, \ldots p_t$, $p_t \leq M_t$ y $l^t_h \neq l^t_q$ si $h \neq q$, la dimensión óptima viene dada por $P = 2M_t - p_t$ y el vector auxiliar óptimo $\mathbf{t_{OP}}$ viene dado por:

$$\mathbf{t_{OP}}(t) = (b^t_1, a^t_1 \ldots, b^t_{j_1-1}, a^t_{j_1-1}, a^t_{j_1}, b^t_{j_1+1}, \ldots, b^t_{j_h-1}, a^t_{j_h-1}, a^t_{j_h}, b^t_{j_h+1}, \ldots b^t_{M_t}, a^t_{M_t}). \tag{3.46}$$

Al igual que anteriormente, el vector $\mathbf{t_{OPT}}(t)$ depende de valores poblacionales de la variable de estudio y, por lo que debemos obtener una estimación $\mathbf{t_{OPTs}}(t)$ del vector óptimo $\mathbf{t_{OPT}}(t)$ a partir de los conjuntos A_{st} y B_{st}. A partir del vector $\mathbf{t_{OPTs}}(t)$ se construye el estimador de Martínez et al. (2017) mediante el proceso de calibración descrito para el estimador $\widehat{F}_{yc}(t)$.

Otro inconveniente asociado al vector óptimo $\mathbf{t_{OPT}}(t)$ es que para cada valor t para el que se quiera obtener una estimación de $F_y(t)$, tenemos una opción diferente de $\mathbf{t_{OPT}}(t)$ y de $\mathbf{t_{OPTs}}(t)$, de forma que el estimador de Martínez et al. (2017) para cada valor t vendría dado por:

$$\widehat{F}_{ycopt}(t) = \frac{1}{N}\sum_{k\in s}\omega_k(t)\Delta(t-y_k) \tag{3.47}$$

donde los pesos calibrados $\omega_k(t)$ se obtienen con el proceso usual de calibración a partir de la estimación $\mathbf{t_{OPTs}}(t)$ y en consecuencia depende del punto t donde se quiere estimar la función de distribución, lo que puede afectar a la propiedad de monotonía no decreciente. En el Apéndice 3 se analizan las condiciones bajo las cuales el estimador calibrado de Martínez et al. (2017) satisface todas las propiedades de función de distribución.

Finalmente, en ocasiones tanto $\mathbf{t_{OPT}}(t)$ como $\mathbf{t_{OPTs}}(t)$ pueden tener una dimensión muy grande lo que puede provocar algunos problemas, como procesos de calibración sin solución o sobrecalibración (Chauvet & Goga, 2022). Una de las consecuencias de la sobrecalibración es la pérdida de eficiencia del estimador calibrado (Chauvet & Goga, 2022). En el Apéndice 4, analizaremos las condiciones en las que la dimensión del vector óptimo $\mathbf{t_{OPT}}(t)$ puede reducirse sin pérdida de eficiencia.

3.5. Estimación de cuantiles y medidas de pobreza en poblaciones finitas a partir de la función de distribución

En esta sección revisaremos cómo estimar cuantiles y medidas de pobreza basadas en cuantiles en poblaciones finitas. A pesar de que existen diversos procedimientos para incorporar la información auxiliar en la estimación de cuantiles, en nuestro caso optaremos por el procedimiento basado en la inversión de un estimador de la función de distribución $F_y(t)$ y revisaremos bajo qué condiciones podemos aplicar dicho procedimiento. Asimismo, analizaremos qué estimadores indirectos de la función de distribución revisados en la sección anterior satisfacen estas condiciones en cuyo caso podrán emplearse directamente en la estimación de cuantiles y en el caso de aquellos estimadores que no cumplan los requisitos, proporcionaremos un método alternativo para poder aplicarlos en la estimación de cuantiles

En el caso de estar interesados en estimar el cuantil de orden α de la variable de estudio y en la población U, asumiremos que una muestra s ha sido seleccionada bajo el diseño muestral $p(\cdot)$. Basándonos en la función de distribución $F_y(t)$ de la variable y, el cuantil $Q_y(\alpha)$ de orden α de la variable y puede definirse de la siguiente manera:

$$Q_y(\alpha) = \text{ínf}\{t : F_y(t) \geq \alpha\} = F_y^{-1}(\alpha) \tag{3.48}$$

Para definir un estimador indirecto del cuantil $Q_y(\alpha)$, podemos incorporar la información auxiliar para obtener un estimador indirecto $\widehat{F}_y(t)$ de la función de distribución $F_y(t)$ de forma que el estimador $\widehat{F}_y(t)$ satisfaga las propiedades de función de distribución. Asumiendo que el estimador $\widehat{F}_y(t)$ es una auténtica función de distribución, un estimador indirecto para el cuantil $Q_y(\alpha)$ vendría dado por:

$$\widehat{Q}_y(\alpha) = \text{ínf}\{t : \widehat{F}_y(t) \geq \alpha\} = \widehat{F}_y^{-1}(\alpha) \tag{3.49}$$

Las propiedades que un estimador $\widehat{F}_y(t)$ debe cumplir para se una auténtica función de distribución son:

i. $\widehat{F}_y(t)$ es continua a la derecha.

ii. $\widehat{F}_y(t)$ es monótono, no decreciente.

iii. a) $\lim_{t\to-\infty} \widehat{F}_y(t) = 0$ and b) $\lim_{t\to+\infty} \widehat{F}_y(t) = 1$.

En el caso de que el estimador considerado $\widehat{F}_y(t)$ no sea monótono no decreciente, se puede aplicar el procedimeinto general descrito en Rao et al. (1990). Para ello, debemos considerar los valores muestrales de la variable y ordenados de menor a mayor, esto es:

$$y_{(1)} < y_{(2)} \ldots y_{(r)} \text{ con } r \leq n$$

A continuación definimos un nuevo estimador $\tilde{F}_y(t)$ de la siguiente manera:

$$\tilde{F}_y(y_{(1)}) = \widehat{F}_y(y_{(1)}) \quad ; \quad \tilde{F}_y(y_{(i)}) = \text{máx}\left\{\tilde{F}_y(y_{(i-1)}), \widehat{F}_y(y_{(i)})\right\} \text{ para } i = 2, \ldots r \tag{3.50}$$

Con ello, el nuevo estimador $\tilde{F}_y(t)$ respeta la propiedad de monotonía no decreciente y puede emplearse en la estimación de cuantiles mediante (3.49).

Un estimador $\widehat{F}_y(t)$ que no satisfaga la condición $\lim_{t\to+\infty} \widehat{F}_y(t) = 1$, también se puede emplear en la

estimación del cuantil $Q_y(\alpha)$ siempre y cuando

$$\text{máx}\{\widehat{F}_{YO}(y_i) : i \in s\} \geq \alpha. \tag{3.51}$$

Entre los estimadores indirectos revisados, el estimador de Háyek $\widehat{F}_{YHJ}(t)$, el estimador de Chambers-Dunstan $\widehat{F}_{CD}(t)$ y el estimador calibrado de Rueda et al. (2007a) $\widehat{F}_{yc}(t)$ son aútenticas funciones de distribución y en consecuencia pueden ser empleadas en la estimación de cuantiles. Por otro lado, el estimador de Horvitz-Thompson, no satisface en general la condición $\lim_{t\to+\infty} \widehat{F}_y(t) = 1$ pero puede ser empleado en la estimación del cuantil $Q_y(\alpha)$ si se satisface la condición (3.33).

Respecto a los estimadores de razón $\widehat{F}_{YR}(t)$, diferencia $\widehat{F}_{YD}(t)$ y Rao-Kovar-Mantel $\widehat{F}_{RKM}(t)$, todos ellos no satisfacen en general la monotonía no decreciente y por ello es necesario considerar el proceso descrito en (3.50) para poder aplicar estos estimadores en la estimación de cuantiles.

Finalmente, respecto al estimador óptimo de Martínez et al. (2017) $\widehat{F}_{ycopt}(t)$, en el Apéndice 3 se demuestra que bajo condiciones leves se trata de un estimador monótono no decreciente y por tanto se puede emplear en la estimación de cuantiles. En el Apéndice 3 también se abordará cómo optimizar este estimador para la estimación de cuantiles.

Como se mencionó anteriormente, la medición de la pobreza, la desigualdad salarial, la desigualdad y las condiciones de vida son temas de gran interés tanto para la investigación económica (Darvas, 2019; Meyer & Sullivan, 2012; Sompolska-Rzechu la & Kurdyś-Kujawska, 2022) como para los gobiernos, instituciones oficiales y sociedad en general (European Commission, 2010a; Eurostat Statistics, 2020; Jones & Weinberg, 2000; Meglio, 2018).

Existen diversas medidas para la correcta medición de la pobreza, pero algunas de ellas se basan en cuantiles o ratios de cuantiles dado que variables como los salarios o ingresos suelen mostrar asimetría, de forma que los cuantiles son medidas más adecuadas que por ejemplo la media. De ahí la importancia de tener estimadores de cuantiles que sean capaces de incorporar información auxiliar pues generalmente los estudios de pobreza suelen incorporar la medición de un gran número de variables adicionales relacionadas con las condiciones de vida de los individuos encuestados, tales como edad, sexo, educación, empleo, etc (INE, 2022).

De entre las medidas de pobreza empleadas, podemos mencionar que Eurostat fija actualmente el umbral de

pobreza (esto es el umbra para clasificar a cada miembro de la población en pobre y no pobre) igual al sesenta por ciento de la mediana $Q_y(0{,}5)$ de la renta neta equivalente (Eurostat Statistics, 2022). De este modo, se pueden aplicar los estimadores de cuantiles anteriormente mencionados para la estimación del umbral de pobreza. En el Apéndice 3 se incluye un estudio de simulación basado en datos reales de la Encuesta de condiciones de vida en España de 2008 realizada por el Instituto Nacional de Estadística (INE, 2015) donde se compara la eficiencia de diferentes estimadores a la hora de estimar el umbral de pobreza.

Por otro lado, como se mencionó anteriormente en la Introducción, las ratios percentiles han sido considerados como medidas de desigualdad salarial tanto por organismos oficiales (Eurostat Products Datasets, 2022; Countouris et al., 2020; Shrider et al., 2021) como por la investigación económica (Burtless, 1999; Dickens & Manning, 2004; Jones & Weinberg, 2000; Machin et al., 2003). Dada una población U, donde tenemos definida una variable de estudio y, y dados dos valores tal que $1 > \alpha_1 > \alpha_2 > 0$ el ratio de percentiles $R(\alpha_1, \alpha_2)$ viene dado por:

$$R(\alpha_1, \alpha_2) = \frac{Q_y(\alpha_1)}{Q_y(\alpha_2)} \tag{3.52}$$

de forma que si se dispone de un estimador de cuantiles $\widehat{Q}_y(\alpha)$, el ratio $R(\alpha_1, \alpha_2)$ puede ser estimado de la siguiente manera:

$$\widehat{R}(\alpha_1, \alpha_2) = \frac{\widehat{Q}_y(\alpha_1)}{\widehat{Q}_y(\alpha_2)} \tag{3.53}$$

De este modo, podemos emplear los estimadores de cuantiles asociados a los estimadores indirectos de la función de distribución anteriormente mencionados en la estimación de ratios de cuantiles.

En el Apéndice 2 se aborda la estimación de ratios de percentiles a través de técnicas de calibración para el tratamiento de falta de respuesta. Adicionalmente, en el Apéndice 2 se incluye un estudio de simulación para comparar el comportamiento de las técnicas propuestas para la estimación de ratios de percentiles. Concretamente, se ha empleado datos reales procedentes de la Encuesta de condiciones de vida en España de 2016 (INE, 2015) donde se compara la eficiencia de diferentes estimadores de ratios de percentiles bajo falta de respuesta.

Adicionalmente, en el Apéndice 3 también se incluye un estudio de simulación para analizar el comportamiento del estimador de ratios de percentiles asociado al estimador de Martínez et al. (2017) frente a otras alternativas.

3.6. Técnicas de calibración para la falta de respuesta

El método calibración que originalmente había sido concebido para tratar de corregir errores de muestreo (Deville & Särndal, 1992), actualmente ha sido considerado como una de las técnicas más atractivas para el ajuste por falta de respuesta (Lundström & Särndal, 1999). En esta sección, revisaremos las principales técnicas de calibración para el tratamiento de falta de respuesta. Principalemente, se han considerado técnicas de ponderación para la falta de respuesta en la estimación de totales y medias (Beaumont, 2005; Chang & Kott, 2008; Kott & Liao, 2012, 2015, 2017; Lesage et al., 2019; Jo et al., 2015) siendo de menor interés, en nuestro conocimiento, el desarrollo de técnicas específicas para el tratamiento de falta de respuesta en el caso de estimar la función de distribución. A continuación se va a proceder a una revisión de las técnicas de tratamiento de falta de respuesta para la estimación de totales y medias que en el Apéndice 1 son adaptadas a la estimación de la función de distribución.

Para ello, consideraremos una población U de tamaño N donde una muestra s ha sido seleccionada a partir del diseño muestra $p(\cdot)$. También asumiremos la presencia de falta de respuesta en la muestra s respecto a la variable de estudio y, esto es, la variable de estudio y sólo ha podido ser observada en un subconjunto de la muestra original s. Con respecto al vector de información auxiliar $\mathbf{x}_k$, asumiremos que su valor es conocido para toda unidad poblacional.
En consecuencia, bajo la suposición de falta de respuesta, la muestra s puede ser dividida en los siguientes conjuntos disjuntos:

$$s_r = \{k \in s/\text{donde la unidad } k \text{ responde}\} \text{ y } s_m = \{k \in s/\text{donde } k \text{ no responde}\},$$

Así s_r representa la muestra de unidades sin falta de respuesta en la variable y, cuyo tamaño denotaremos por r y s_m que representa la muestra de unidades con falta de respuesta en la variable y cuyo tamaño viene dado por $n - r$.

Dado que sólo disponemos de los valores de la variable de estudio y en la muestra s_r, para estimar el total poblacional T_y, se podría considerar el estimador:

$$\widehat{T}_{YH} = \sum_{k \in s_r} d_k \Delta(t - y_k)$$

El estimador $\widehat{T}_{YH}$ puede proporcionar estimaciones sesgadas debido a la falta de representación de ciertos

grupos específicos. El sesgo producido por el estimador $\widehat{T}_{YH}$ puede tratar de corregirse mediante el uso de la reponderación. Mediante la ponderación, se determina un conjunto de pesos por medio de la incorporación de la información auxiliar disponible, y se realiza la estimación aplicando los pesos a los valores de y para los elementos que respondieron, siendo la calibración unos de los métodos empleados para reponderar (Särndal & Lundström, 2005).

Bajo este enfoque, el conjunto de respuesta s_r se obtiene como un subconjunto de s y supondremos que los elementos incluidos en la muestra original s responden de forma independiente y que la distribución de la respuesta tiene probabilidades de respuesta de primer orden $P(k \in s_r|s) = p_k \geq 0$.

Dado que estas probabilidades de respuesta son desconocidas, debemos considerar estimaciones $\widehat{p}_k$ para poder obtener un estimador para el total poblacional de y bajo falta de respuesta, reemplazando los pesos de diseño originales d_k por $d_k^o = (\pi_k \hat{p}_k)^{-1}$, de forma que el estimador del total vendría dado por:

$$\widehat{T}_{Yw}(t) = \sum_{k \in s_r} d_k^o y_k$$

Dado que el estimador $\widehat{T}_{Yw}(t)$ no incorpora la información auxiliar, la idea es susitutir los pesos d_k^o por unas nuevas reponderaciones que incorporen la información auxiliar disponible para corregir el sesgo provocado por la falta de respuesta. A continuación, bajo este enfoque revisaremos algunas técnicas de calibración empleadas en la estimación de totales y medias bajo falta de respuesta.

3.6.1. Estimador de Lundström & Särndal (1999)

Bajo la presencia de falta de respuesta, Lundström & Särndal (1999) propusieron emplear el método de calibración en la estimación de un total poblacional. Concretamente, el estimador de Lundström & Särndal (1999) considera la minimización de (3.18) para la muestra s_r, esto es:

$$\sum_{k \in s_r} \frac{(\omega_k - d_k)^2}{d_k q_k}$$

bajo la siguiente restricción:

$$\sum_{k \in s_r} \omega_k \mathbf{x}_k = T_{\mathbf{x}} \tag{3.54}$$

Los pesos calibrados así obtenidos viene dados por $\omega_k = d_k \upsilon_k$, donde:

$$\upsilon_k = 1 + q_k \Big(T_{\mathbf{x}} - \sum_{k \in s_r} \mathbf{x}_k \Big)' \cdot \Big(\sum_{k \in s_r} d_k q_k \mathbf{x}_k \mathbf{x}_k' \Big)^{-1}$$

obteniendo el estimador

$$\widehat{T}_{LS} = \sum_{k \in s_r} \omega_k y_k \tag{3.55}$$

Lundström & Särndal (1999) establecieron que el error cuadrático medio del estimador $\widehat{T}_{LS}$ viene dado por:

$$MSE(\widehat{T}_{LS}) = V_p + V_{nr} + 2Cov_p(\widehat{T}_{HT}), B_{nr|s}) + E_p[B^2_{nr|s}] \tag{3.56}$$

donde V_p denota la varianza bajo el diseño $p(\cdot)$ del estimador de Horvitz-Thompson $\widehat{T}_{HT}$ para el total y, esto es $V_p = V_p(\widehat{T}_{HT})$, V_{nr} denota la varianza del error por falta de respuesta dado por $E_p[V_q(\widehat{T}_{LS}|s)]$ donde V_q denota la varianza bajo la distribución de respuesta asumida y $B_{nr|s} = E_q[\widehat{T}_{LS} - \widehat{T}_{HT}|s]$ es el sesgo de no respuesta (condicionado a s).

Bajo la suposición de que $B_{nr|s} = 0$, y de que $p_{kl} = P[k, l \in s_r|s] = p_k p_l$ para todo $k \neq l$, Lundström & Särndal (1999) establecieron un estimador para (3.56) (para más detalles consultar Lundström & Särndal (1999)).

3.6.2. Estimador de Deville (2000)

Bajo el enfoque de Deville (2000) consideramos un modelo más flexible donde el método de calibración considerado distinguirá entre variables que modelan el mecanismo de respuesta $\mathbf{x}^*_k$ de dimensión M y variables que intervienen en las restricciones de calibración $\mathbf{z}_k$ de dimension J, de forma que supondremos conocidos tanto el total poblacional $T_{\mathbf{x}^*}$ como el total poblacional $T_{\mathbf{z}}$. Adicionalmente, asumiremos como suposición la condición de datos faltantes al azar (missing at random MAR), esto es, la variable de estudio y no influye en el mecanismo de respuesta. Bajo este supuesto, la probabilidad de respuesta puede ser modelada de la siguiente manera:

$$p_k = f(\gamma^{'} \mathbf{x}^*_k)$$

donde γ es un vector de parámetros, $h(\cdot) = 1/f(\cdot)$ es una función conocida, monótona y dos veces diferenciable.

Ejemplos de modelos de respuestas de este tipo pueden ser (Kott & Liao, 2012):

- El modelo lineal, dado por:

$$p_k = 1 + \gamma^{'} \mathbf{x}^*_k$$

- el modelo raking, dado por:

$$p_k = \frac{1}{exp(-\gamma^{'} \mathbf{x}^*_k)}$$

- El modelo logístico (l, u) dado por:

$$p_k = \frac{1 + exp(\gamma' \mathbf{x}_k)^*/u}{l + exp(\gamma' \mathbf{x}_k)^*}$$

- El modelo logístico, dado por:

$$p_k = \frac{exp(\gamma' \mathbf{x}_k^*)}{1 + exp(\gamma' \mathbf{x}_k^*)}$$

En particular, el modelo logístico es un caso particular del modelo logit (l, u) donde $u = \infty$, $c = 2$ y $l = 1$ (Kott & Liao, 2012).

Bajo este enfoque, vamos a considerar que los pesos calibrados tiene la siguiente forma

$$\omega_k = \frac{d_k}{f(\hat{\gamma}' \mathbf{x}_k^*)} = d_k h(\hat{\gamma}' \mathbf{x}_k^*) \tag{3.57}$$

de forma que satisfagan la siguiente restricción:

$$\sum_{s_r} \frac{d_k}{f(\hat{\gamma}' \mathbf{x}_k^*)} \mathbf{z}_k = \sum_{s_r} d_k h(\hat{\gamma}' \mathbf{x}_k^*) \mathbf{z}_k = T_{\mathbf{z}} \tag{3.58}$$

donde $\hat{\gamma}$ es un estimador consistente de γ. El estimador resultante para el total poblacional Y viene dado por:

$$\widehat{T}_{YDC} = \sum_{s_r} d_k \frac{1}{\hat{p}_k} y_k = \sum_{s_r} d_k h(\hat{\gamma}' \mathbf{x}_k^*) y_k \tag{3.59}$$

En la propuesta de Deville (2000) las dimensiones M y J deben ser iguales, lo que no implica que los vectores de variables $\mathbf{x}_k^*$ y $\mathbf{z}_k$ necesariamente coincidan. De hecho, los vectores $\mathbf{x}_k^*$ y $\mathbf{z}_k$ pueden tanto tener alguna variable en común como no coincidir en ninguna de sus variables. Para mayor detalle sobre este estimador, puede consultarse Deville (2000).

3.6.3. Estimador de Kott & Liao (2017)

Kott & Liao (2017) extendieron la propuestas de Deville (2000) al caso donde se permiten más variables de calibración que variables para la modelización de falta de respuesta, esto es, el caso donde $J > M$. Para ello, asumieron el marco conceptual propuesto en Chang & Kott (2008), donde en lugar de buscar un estimador $\widehat{\gamma}$ que satisfaga la ecuación (3.58, lo que se pretende, dada una matriz Q simétrica y definida positiva es buscar un estimador $\widehat{\gamma}$ que minimice:

$$\mathbf{v}' Q \mathbf{v} \tag{3.60}$$

donde

$$\mathbf{v} = \frac{1}{N}\left(\sum_{s_r} d_k h(\hat{\gamma}^T \mathbf{x}_k^*)\mathbf{z}_k - \sum_s d_k \mathbf{z}_k\right)$$

La minimización de (3.60) es equivalente a reformular el siguiente proceso de calibración:

$$\sum_{s_r} \omega_k \tilde{\mathbf{z}}_k = \sum_{s_r} d_k h(\hat{\gamma}^{'} \mathbf{x}_k^*)\tilde{\mathbf{z}}_k = \sum_s d_k \tilde{\mathbf{z}}_k \tag{3.61}$$

con $\tilde{\mathbf{z}}_k = A \cdot \mathbf{z}_k$ donde

$$A = \left[\frac{1}{N}\sum_{s_r} d_i h^{'}(\hat{\gamma}^{'} \mathbf{x}_i^*) x_i^* \mathbf{z}_i^{'} Q\right]$$

donde $h^{'}(\cdot)$ denota la primera derivada de $h(\cdot)$.

Para poder llevar a cabo todo este proceso, necesitamos seleccionar una matriz Q simétrica y definida positiva. La literatura previa ha considerado diversas formas de seleccionar la matriz Q (Kott & Liao, 2017). Una elección obvia es seleccionar la matriz Q igual a la identidad. Una elección más apropiada puede ser la siguiente:

$$Q^{-1} = diag\left(\frac{1}{N}\sum_{k\in s} d_k \mathbf{z}_k\right)$$

Con esta opción el estimador calibrado obtenido es invariante a los cambios de escala de medida en el vector $\mathbf{z}_k$.

Por otro lado, Chang & Kott (2008) consideraron como elección para la matriz Q una estimación de:

$$\tau = \frac{1}{N}\left(\sum_{k\in s_r} \frac{d_k}{f(\gamma^{'} \mathbf{x}_k^*)} \cdot \mathbf{z}_k\right)$$

El problema de la anterior elección, es que se requiere un método iterativo para determinar el valor de Q ya que la estimación empleada de γ influye en la estimación de Q que a su vez también influye en la estimación de γ. Para evitar este problema, Kott & Liao (2017) propusieron dos elecciones de la matriz Q alternativas.

En la primera de ellas, la matriz Q viene dada por

$$Q = \left(\frac{1}{N}\sum_{s_r} d_k h^{'}(\hat{\gamma}^{'} \mathbf{x}_k^*)\mathbf{z}_k (\mathbf{z}_k)^{'}\right)^{-1} \tag{3.62}$$

De este modo, la estimación de γ y Q se obtienen mediante un proceso iterativo para satisfacer la ecuación

(3.61). Con ello, se obtiene el siguiente estimador calibrado:

$$\widehat{T}^{1}_{YKL} = \sum_{k \in s_r} \omega^{1}_{k} y_k = \sum_{s_r} d_k h(\hat{\gamma}^{'} \mathbf{x}^{*}_{k}) y_k \tag{3.63}$$

En la segunda propuesta, Kott & Liao (2017) en lugar de emplear una matriz Q para el proceso de calibración dado por la minimización de $\mathbf{v}^{'} Q \mathbf{v}$, basándose en la técnica de reducción de dimension propuesto en Andridge & Little (2011), tratan de buscar unos pesos calibrados que satisfagan la restricción (3.61) donde $\tilde{\mathbf{z}}_k = A_0 \cdot \mathbf{z}_k$ de forma que A_0 viene dado por:

$$A^{T}_{0} = \left(\sum_{s_r} \mathbf{z}_j (\mathbf{z}_j)^{'} \right)^{-1} \sum_{s_r} \mathbf{z}_j (\mathbf{x}^{*}_{j})^{'}$$

Con los nuevos pesos calibrados, se obtiene el siguiente estimador:

$$\widehat{T}^{2}_{YKL} = \sum_{k \in s_r} \omega^{2}_{k} y_k = \sum_{s_r} d_k h(\hat{\gamma}^{'} \mathbf{x}^{*}_{k}) y_k \tag{3.64}$$

Para un estudio más detallado de estos dos estimadores puede consultarse Kott & Liao (2017).

En el Apéndice 1 se aborda la extensión de las técnicas para la estimación de totales y medias propuestas tanto por Deville (2000) como por Kott & Liao (2017) a la estimación de la función de distribución y se propondrán nuevos estimadores de la función de distribución bajo la presencia de falta de respuesta. De igual forma, en el Apéndice 1 también se abordan las condiciones bajos las cuales los nuevos estimadores propuestos son auténticas funciones de distribución.

3.7. Técnicas de remuestreo para la estimación de la varianza

Dado que tanto los cuantiles como los ratios de percentiles no son parámetros lineales, en ocasiones no será posible establecer una expresión de la varianza de los estimadores propuestos, debido a su complejidad. Por ello, para poder realizar estimaciones de la varianza de los estimadores propuestos en ocasiones debemos recurrir a técnicas bootstrap que nos permitirán tanto estimar la varianza como obtener intervalos de confianza para los estimadores calibrados desarrollados en la presente tesis. Concretamente, las técnicas bootstrap consideradas son las propuestas por Booth et al. (1994)y Antal & Tillé (2011, 2014).

A continuación, pasamos a realizar una breve revisión de todas ellas. Para ello, ilustraremos cómo aplicar las técnicas bootstrap en el caso de estimadores para cuantiles, siendo su desarrollo análogo para el caso de

ratios de percentiles.

3.7.1. Técnica bootstrap de Booth et al. (1994)

Si disponemos de un estimador para cuantiles $\widehat{Q}_y(\alpha)$, la técnica propuesta por Booth et al. (1994) se basa en obtener poblaciones artificiales a partir de copias de la muestra s seleccionada. Para ello, Booth et al. (1994) considera que si el tamaño poblacional N viene dado por $N = n \cdot c + a$ con $0 < a < n$, podemos construir una población artificial U_B mediante c copias de la muestra s a la que le añadimos una muestra de tamaño a obtenidad a partir de la muestra s mediante muestreo aleatorio simple sin reemplazamiento. Mediante este proceso, se pueden obtener M poblaciones artificiales distintas U_B^j, $j = 1, \ldots, M$ y para cada una de ellas, podemos seleccionar K muestras bootstrap $s_1^j, \ldots, s_K^j$ de tamaño n. Con la muestra s_h^j, podemos calcular la estimación $\widehat{Q}_y^*(\alpha)_h^j$ para la población U_B^j de forma que la estimación de la varianza viene dada por (Chauvet, 2007):

$$\widehat{V}(\widehat{Q}_y(\alpha)) = \frac{1}{M}\sum_{j=1}^{M}.\widehat{V}_j. \tag{3.65}$$

donde

$$\widehat{V}_j = \frac{1}{K-1}\sum_{h=1}^{K}(\widehat{Q}_y^*(\alpha)_h^j - \widehat{Q}_y^*(\alpha)^j)^2$$

$$\widehat{Q}_y^*(\alpha)^j = \frac{1}{K}\sum_{h=1}^{K}\widehat{Q}_y^*(\alpha)_h^j$$

3.7.2. Técnicas bootstrap de Antal & Tillé (2011) y Antal & Tillé (2014)

A diferencia de Booth et al. (1994) que considera muestras bootstrap a partir de poblaciones artificiales, Antal & Tillé (2011) y Antal & Tillé (2014) han propuesto recientemente técnicas bootstrap directas donde las muestras bootstrap son seleccionadas a partir de la muestra s original sin necesidad de construir poblaciones artificiales. Para ello, las muestras bootstrap son obtenidas mediante diseños muestrales diferentes al diseño muestral originalmente considerado para seleccionar la muestra s.

Concretamente, si se ha seleccionado la muestra de partida s mediante muestreo aleatorio simple, Antal & Tillé (2011) obtienen dos muestras con dos diseños diferentes a partir de la muestra s. Por un lado, obtiene una muestra de s a partir de muestreo aleatorio simple sin reemplazamiento y por otro lado obtiene otra muestra de s a partir de un diseño definido por los autores para el remuestreo llamado one-one sampling.

Similarmente, Antal & Tillé (2014) proponen un diseño muestral basado también en dos diseños muestrales. En su caso, a partir de la muestra s se obtiene una primera muestra con diseño de Bernoulli y una segunda muestra a partir de un diseño establecido por los autores y denominado double half sampling. Para una mayor

revisión de ambas técnicas bootstrap puede consultarse Antal & Tillé (2011) y Antal & Tillé (2014).

Para ambos enfoques, si consideramos un estimador de cuantiles $\widehat{Q}_y(\alpha)$ y una muestra s, M muestras bootstrap $s_1^*, \ldots, s_M^*$ son seleccionadas de acuerdo al esquema de muestreo descrito tanto en Antal & Tillé (2011) como en Antal & Tillé (2014). La estimación de la varianza para el estimador $\widehat{Q}_y(\alpha)$ es:

$$\widehat{V}(\widehat{Q}_y(\alpha)) = \frac{1}{M}\sum_{j=1}^{M}(\widehat{Q}_y(\alpha)_j^* - \bar{Q}_y(\alpha)^*)^2 \tag{3.66}$$

donde

$$\bar{Q}_y(\alpha)^* = \frac{1}{M}\sum_{j=1}^{M}\widehat{Q}_y(\alpha)_j^*.$$

y $\widehat{Q}_y(\alpha)_j^*$ denota el estimador bootstrap obtenido a partir de la muestra bootstrap s_j^*.

Capítulo 4

Resultados

La investigación llevada a cabo para el desarrollo de la presente tesis doctoral ha originado algunos resultados relevantes. De entre los resultados alcanzados, a continuación pasamos a detallar los más relevantes.

4.1. Treating nonresponse in the estimation of the distribution function

Respecto al tratamiento de la falta de respuesta mediante técnicas de calibración en la estimación de la función de distribución, los principales resultados alcanzados son:

- Se proponen cinco estimadores de la función de distribución bajo falta de respuesta, $\hat{F}_{cal}(t)$, $\hat{F}_{calTS}(t)$, $\widehat{F}_{calID}$, $\widehat{F}_{calIKL1}$ y $\widehat{F}_{calIKL2}$. Los estimadores $\hat{F}_{cal}(t)$ y $\hat{F}_{calTS}(t)$ adaptan la metodología propuesta en Rueda et al. (2007a) para el tratamiento de falta de respuesta, si bien el estimador $\hat{F}_{calTS}(t)$ lo hace mediante calibración en dos pasos (Kott & Liao, 2015). Los estimadores $\widehat{F}_{calID}$, $\widehat{F}_{calIKL1}$ y $\widehat{F}_{calIKL2}$ extienden la calibración con variables modelizadoras de la falta de respuesta e instrumentales propuestas en Deville (2000) y Kott & Liao (2017).

- El estimador $\hat{F}_{calTS}(t)$ cumple en general todas las propiedades de función de distribución excepto la monotonía no decreciente y la unicidad del límite en $+\infty$. Si bien, esta última propiedad puede ser satisfecha si en las restricciones de calibración, el último punto considerado t_p es lo suficientemente grande para que $F_{\tilde{y}}(t_P) = 1$.

- El estimador $\widehat{F}_{calID}$ basado en la propuesta de Deville (2000) cumple todas las propiedades de función de distribución excepto $\lim_{t \to +\infty} \widehat{F}_y(t) = 1$, pero al igual que en el caso anterior se puede garantizar con un punto t_P tal que $F_{\tilde{y}}(t_P) = 1$.

- Los estimadores $\widehat{F}_{calIKL1}$ y $\widehat{F}_{calIKL2}$ basados en la propuesta de Kott & Liao (2017) tampoco

satisfacen en general la unicidad del límite en $+\infty$. Mediante un teorema se establece que si el vector x_k^* contiene una variable constantemente igual a 1 y se considera el valor t_P suficientemente grande, los estimadores $\widehat{F}_{calIKL1}$ y $\widehat{F}_{calIKL2}$ satisfacen la condición $\lim_{t\to+\infty} \widehat{F}_y(t) = 1$.

- Mediante estudios de simulación se pone de manifiesto que los estimadores $\hat{F}_{calTS}(t)$ y $\widehat{F}_{calIKL1}$ y $\widehat{F}_{calIKL2}$ muestran el mejor comportamiento en términos de eficiencia.

4.2. Calibration adjustment for dealing with nonresponse in the estimation of poverty measures

Una vez desarrolladas técnicas de estimación de la función de distribución en presencia de datos faltantes y establecidas las condiciones bajo las cuales los estimadores propuestos son auténticas funciones de distribución, se procede a emplear estas técnicas en la estimación de cuantiles y medidas de pobreza. Los principales resultados obtenidos en este ámbito son:

- Se proponen cinco estimadores de cuantiles $\widehat{Q}_{cal}^{(1)}(\alpha), \widehat{Q}_{cal}^{(2)}(\alpha), \widehat{Q}_{D}^{(3)}(\alpha), \widehat{Q}_{KL1}^{(3)}(\alpha)$ y $\widehat{Q}_{KL2}^{(3)}(\alpha)$ basados respectivamente en los estimadores $\hat{F}_{cal}(t)$, $\hat{F}_{calTS}(t)$, $\widehat{F}_{calID}$, $\widehat{F}_{calIKL1}$ y $\widehat{F}_{calIKL2}$ obtenidos en el Apéndice 1. Para ello, se consideran las condiciones establecidas en el Apéndice 1 bajo las cuales los estimadores de la función de distribución son auténticas funciones de distribución.

- Los correspondientes cinco estimadores de ratios de cuantiles $\widehat{R}_{cal}^{(1)}(\alpha_1, \alpha_2), \widehat{R}_{cal}^{(2)}(\alpha_1, \alpha_2), \widehat{R}_{D}^{(3)}(\alpha_1, \alpha_2)$, $\widehat{R}_{KL1}^{(3)}(\alpha_1, \alpha_2)$ y $\widehat{R}_{KL2}^{(3)}(\alpha_1, \alpha_2)$ también son definidos.

- Para poder establecer los estimadores $\widehat{Q}_{cal}^{(2)}(\alpha)$ y $\widehat{R}_{cal}^{(2)}(\alpha_1, \alpha_2)$, dado que el estimador $\hat{F}_{calTS}(t)$ no es en geneal monótono no decreciente, es necesario aplicar el procedimiento descrito en (3.50).

- Para todos los estimadores de ratios de percentiles propuestos se proporcionan estimadores bootstrap de sus respectivas varianzas mediante la aplicación de las técnicas descritas en Antal & Tillé (2011, 2014) y Booth et al. (1994).

- Los resultados de un estudio de simulación con datos reales procedentes de la Encuesta de condiciones de vida (INE, 2015) correspondientes al año 2008 ponen de manifiesto que los estimadores propuestos son los que mejor comportamiento presentan en términos de eficiencia. Adicionalmente, respecto a la estimación de varianza, los intervalos de confianza asociados a los estimadores propuestos alcanzan una alta cobertura con menor amplitud que otras alternativas consideradas.

4.3. The optimization problem of quantile and poverty measures estimation based on calibration

Respecto al análisis de las condiciones bajo las cuales el estimador propuesto en Martínez et al. (2017) para la función de distribución satisface las propiedades de función de distribución, los resultados más notables son:

- El estimador de Martínez et al. (2017) cumplen en general todas las propiedades de función de distribución excepto la monotonía no decreciente y $\lim_{t\to+\infty} \widehat{F}_y(t) = 1$.
- Se establece un teorema que demuestra que si en el proceso de calibración las constantes $q_k = 1$ para toda unidad poblacional, el estimador de Martínez et al. (2017) es monótono no decreciente.
- La condición $\lim_{t\to+\infty} \widehat{F}_y(t) = 1$ puede ser alcanzada considerando un punto t_P suficientemente grande, pero añadir esta condición supone que el estimador ya no minimizaría la varianza asintótica y por tanto no sería optimo.
- Se establecen los estimadores de cuantiles y de ratios de cuantiles asociados al estimador de Martínez et al. (2017), pero dado que no se verifica la propiedad $\lim_{t\to+\infty} \widehat{F}_y(t) = 1$, es necesario que se cumpla la condición (3.33).
- Tanto para los estimadores de cuantiles como para los estimadores de ratios de cuantiles se proporcionan estimadores bootstrap de sus respectivas varianzas mediante la aplicación de las técnicas descritas en Antal & Tillé (2011, 2014) y Booth et al. (1994).
- Un estudio de simulación llevado a cabo con datos reales correspondientes al año 2008 y procedentes de la Encuesta de condiciones de vida (INE, 2015) son empleados para la estimación del umbral de pobreza. A partir de los resultados se observa que que el estimador propuesto es el más eficiente frente a las alternativas consideradas.
- Los resultados de un estudio de simulación llevado a cabo con datos reales correspondientes al año 2016 son empleados para la estimación de los ratios de percentiles y de igual forma, el estimador propuesto presenta la mejor eficiencia.
- Respecto a la estimación de la varianza incluida en ambos estudios de simulación, se concluye que los métodos bootstrap tienden a sobrestimar la varianza de los estimadores para el umbral de pobreza mientras que para la varianza de los estimadores de ratios de percentiles suelen infraestimar la varianza.

4.4. Reduction of optimal calibration dimension with a new optimal auxiliary vector for calibrated estimators of the distribution function

En relación al análisis de las condiciones bajo las cuales la alta dimensionalidad que en ocasiones puede existir en las condiciones de calibración asociadas al estimador propuesto por Martínez et al. (2017) pueden ser reducidas, los resultados alcanzados son:

- Se establecen condiciones bajo las cuales de forma general, el vector óptimo $\mathbf{t_{opt}}$ puede ver reducida su dimensión mediante un vector óptimo alternativo $\mathbf{t_{optred}}$.
- Se propone un nuevo estimador de la función de distribución basado en el nuevo vector óptimo alternativo de menor dimensión.
- A partir de los estudios de simulación realizados, el nuevo vector óptimo de dimensión reducidad $\mathbf{t_{optred}}$ permite reducir de manera bastante considerable el coste computacional a la hora de obtener las estimaciones sin pérdida de eficiencia respecto al vector óptimo original $\mathbf{t_{opt}}$ propuesto en Martínez et al. (2017) y en ocasiones incluso puede mejorar la eficiencia.

Capítulo 5

Conclusiones

5.1. Treating nonresponse in the estimation of the distribution function

En esta aportación se describe como mediante repoderación calibrada se pueden ajustar los pesos básicos del diseño considerado para el tratamiento de falta de respuesta a la hora de estimar la función de distribución. De este modo, se proponen cinco estimadores basados en dos metodologías distintas para reducir el sesgo producido por la falta de respuesta. La primera metodología trata de adaptar la propuesta de Rueda et al. (2007a) para el tratamiento de falta de respuesta y a partir de ella se proponen dos estimadores $\hat{F}_{cal}(t)$ y $\hat{F}_{calTS}(t)$. Para el desarrollo del estimador $\hat{F}_{calTS}(t)$ hemos considerado calibración en dos pasos (Kott & Liao, 2015), donde el primer proceso de calibración está destinado a eliminar el sesgo del falta de respuesta mientras que el segundo de ellos está destinado a reducir el error de muestreo y dado que el modelo de falta de respuesta y el modelo predictivo pueden ser muy diferentes, se permite el empleo de diferentes variables en cada una de las fases. El estimador así obtenido $\hat{F}_{calTS}(t)$ destaca por sus simplicidad computacional.
La segunda metodología se basa en calibración generalizada (Deville, 2000; Kott & Liao, 2017), de forma que la calibración se lleva a cabo en una única etapa pero en la que se permite emplear diferentes conjuntos de variables para modelar la falta de respuesta y para la restricciones de calibración. Esta segunda metodología proporciona tres nuevos estimadores $\widehat{F}_{calID}$, $\widehat{F}_{calIKL1}$ y $\widehat{F}_{calIKL2}$ que requieren de métodos iterativos para resolver las restricciones de calibración y por ello su principal incoveniente es la complejidad computacional frente a la sencillez del estimador $\hat{F}_{calTS}(t)$.
Nuestro estudio de simulación pone de manifiesto claramente la reducción del sesgo y la precisión que se logra cuando se usa la calibración para la falta de respuesta. A pesar de que no existe un estimador que sea uniformemente mejor que el resto en términos de sesgo y eficiencia, los estimadores $\widehat{F}_{calIKL1}$ y $\widehat{F}_{calIKL2}$ producen las mejores estimaciones en términos del menor error en la mayoría de los casos.

5.2. Calibration adjustment for dealing with nonresponse in the estimation of poverty measures

En esta propuesta se proponen técnicas de calibración para la estimación de medidas de pobreza basadas en ratios de percentiles cuando se produce falta de respuesta. El estudio de simulación llevado a cabo con datos reales pone de manifiesto la mejora tanto en términos de sesgo como en términos de eficiencia que se alcanza con dos de los estimadores propuestos $\widehat{R}^{(2)}_{cal}$ y $\widehat{R}^{(3)}_{cali}$. Mientras el estimador $\widehat{R}^{(2)}_{cal}$ se basa en calibración en dos pasos (Kott & Liao, 2015), el estimador $\widehat{R}^{(3)}_{cali}$ se basa en calibración generalizada (Kott & Liao, 2017).

El estudio de simulación abarca una amplia variedad de ratios de percentiles así como de diferentes modelos para la falta de respuesta (lineal, raking y logit) y muestran que en todos los casos se produce un gran descenso en el sesgo y en el error cuadrático medio de los estimadores, lo que muestra la robustez de las técnicas propuestas para el ajuste de falta de respuesta.

A pesar de que los resultados del estudio de simulación no ponen de manifiesto que exista un estimador que sea uniformemente mejor que el resto entre los estimadores propuestos los estimadores $\widehat{R}^{(2)}_{cal}$ y $\widehat{R}^{(3)}_{KL2cal}$ son más simples computacionalmente que el resto y por tanto son alternativas adecuadas a la hora de estimar medidas para la desigualdad salarial basadas en ratios de percentiles.

5.3. The optimization problem of quantile and poverty measures estimation based on calibration

En este trabajo se proporciona un estimador óptimo para la estimación de cuantiles en el sentido de mínima varianza. El estimador propuesto se basa en el estimador calibrado de la función de distribución propuesto por Rueda et al. (2007a) y en el vector óptimo de calibrción para dicho estimador proporcionado en Martínez et al. (2017).

Para ello el punto de partida de esta investigación consiste en formular el problema de minimizar la varianza del estimador de cuantiles y establecer la equivalencia de este problema de minimización con el problema de minimización de la varianza del estimador de la función de distribución asociado. De este modo, partiendo del estimador propuesto por Rueda et al. (2007a) y del vector óptimo propuestoen Martínez et al. (2017), demostramos mediante el establecimiento de un teorema que el estimador calibrado óptimo para la

función de distribución $\widehat{F}_{YO}(t)$ cumple todas las propiedades de función de distribución y no es necesario corregir la falta de monotonía no decreciente mediante la técnica propuesta en Rao et al. (1990).

El estudio de simulación llevado a cabo se desarrolla con datos reales para la estimación del umbral de pobreza y ratios de percentiles para la desigualdad salarial. Los resultados de este estudio muestran que los estimadores para los ratios de percentiles y umbrales de pobreza derivados del estimador óptimo $\widehat{F}_{YO}(t)$ son alternativas adecuadas para la estimación de estas medidas y además pone de manifiesto que los estimadores bootstrap para estimar la varianza de los estimadores de medidas de pobreza no proporcionan estimaciones insesgadas de la misma siendo necesario rescalarlos para alcanzar la insesgadez.

5.4. Reduction of optimal calibration dimension with a new optimal auxiliary vector for calibrated estimators of the distribution function

En esta propuesta, se analiza si el estimador óptimo propuesto por Martínez et al. (2017) para la estimación de la función de distribución basado en el método de calibración puede ser mejorado mediante una reducción de la dimensión óptima del vector óptimo de calibración empleado $\mathbf{t_{opt}}$, lo que reduciría el número de restricciones de calibración.

Considerar un número reducido de restricciones de calibración permite alcanzar algunos beneficios tales como simplificar problemas numéricos derivados del proceso de optimización así como restringir la posibilidad de obtener pesos negativos o de elevado valor que provoquen estimaciones inestables.

Para ello, se establecen de forma teórica las condiciones bajo las cuales la dimensión del vector óptimo propuesto en Martínez et al. (2017) puede ser reducida. De este modo, se proporciona un nuevo vector óptimo alternativo $\mathbf{t_{optred}}$ de menor dimensión y en base al mismo se define un nuevo estimador para la función de distribución $\widehat{F}_{CALNEWOPT}(t)$. Con este nuevo estimador se pretende evitar los problemas derivados de la lata dimensionalidad de la información auxiliar.

El estudio de simulación realizado muestra claramente que con el nuevo estimador se alcanza una mejora bastante considerable en términos de tiempo de ejecución a la hora de obtener las estimaciones sin pérdida de eficiencia o incluso mejorándola, por lo que el nuevo estimdor supone una alternativa especialmente adecuada cuando se afrontan estimaciones de la función de distribución en poblaciones de gran tamaño o cuando el volumen de información auxiliar también es elevado.

Capítulo 6

Futuras líneas de investigación

Como toda investigación, la presente tesis doctoral presenta una serie de limitaciones que requiere una mayor investigación y que brindan futuras líneas de investigación que pueden ser consideradas como una extensión de la presente investigación. De este modo, es necesario enumerar las cuestiones que no han sido resueltas en esta tesis doctoral, estando actualmente algunas de ellas bajo investigación. Concretamente:

- Para la modelización de la falta de respuesta hemos empleado métodos paramétricos pero otras alternativas puede ser empleadas. Así, una futura línea de investigación puede explorar el uso de técnicas de machine learning como regresión spline, redes neuronales o bootsting en la modelización de la falta de respuesta.

- Relacionado con lo anterior, una futura línea de investigación puede considerar la combinación de técnicas de calibración junto con técnicas como Propensity Score Adjustment con el objetivo de reducir el sesgo de falta de respuesta.

- Dada la relevancia de los estudios de pobreza y desigualdad salarial, otra línea de investigación abierta es tratar de aplicar las técnicas de calibración para el tratamiento de falta de respuesta en la estimación de otras medidas de pobreza descritas en Morales et al. (2018) o como la tasa de pobreza (Martínez et al., 2020; Muñoz et al., 2015).

- Dado que el comportamiento de los estimadores calibrados propuestos para el tratamiento de falta de respuesta también dependen de la elección de un vector de puntos auxiliares, futuras investigación pueden abordar el análisis de la elección óptima de dicho vector respecto a la minimización de la varianza.

- Respecto al estimador óptimo de la función de distribución, cuantiles y medidas de pobreza basados

en el vector óptimo propuesto en Martínez et al. (2017), sólo es válido para muestreo aleatorio simple y por tanto es necesario tratar de extender estas técnicas a otros tipos de muestreo.

- De igual forma, la reducción del vector óptimo de calibración obtenida en esta tesis doctoral también queda restringida a muestreo aleatorio simple, siendo necesario investigar cómo extenderla a diferentes tipos de muestreo.

- Adicionalmente, tanto la propuesta de Martínez et al. (2017) como la reducción obtenida en esta tesis doctoral está restringida al empleo de una pseudo-variable g_k que asume un modelo lineal entre la variable de estudio y y las variables auxiliares incluidas en el vector $\mathbf{x}_k$. Por ello, es necesario investigar la selección óptima bajo modelos no lineales.

- Una futura línea de investigación debe analizar las condiciones bajo las cuales el estimador de la función de distribución asociado al vector óptimo de dimensión reducida, es una auténtica función de distribución para así poder aplicarlo en la estimación de cuantiles y medidas de pobreza.

- Finalmente, otras técnicas de reducción de la dimensión deben ser exploradas a la hora de aplicar técnicas de calibración para tratar de resolver la alta dimensionalidad del vector óptimo de calibración.

Bibliografía

Acal, C., Ruiz-Castro, J. E., Aguilera, A. M., Jiménez-Molinos, F., & Roldán, J. B. (2019). Phase-type distributions for studying variability in resistive memories. *Journal of Computational and Applied Mathematics*, *345*, 23–32.

Alba-Fernández, M., Batsidis, A., Jiménez-Gamero, M.-D., & Jodrá, P. (2017). A class of tests for the two-sample problem for count data. *Journal of Computational and Applied Mathematics*, *318*, 220–229.

Andersson, P. G., & Särndal, C.-E. (2016). Calibration for nonresponse treatment: In one or two stepsf. *Statistical Journal of the IAOS*, *32*(3), 375–381.

Andridge, R. R., & Little, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, *27*(2), 153.

Antal, E., & Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, *106*(494), 534–543.

Antal, E., & Tillé, Y. (2014). A new resampling method for sampling designs without replacement: the doubled half bootstrap. *Computational Statistics*, *29*(5), 1345–1363.

Arcos, A., Contreras, J. M., & Rueda, M. M. (2014). A novel calibration estimator in social surveys. *Sociological methods & research*, *43*(3), 465–489.

Arcos, A., Martínez, S., Rueda, M., & Martínez, H. (2017). Distribution function estimates from dual frame context. *Journal of Computational and Applied Mathematics*, *318*, 242–252.

Arcos, A., Rueda, M., & Muñoz, J. F. (2007). An improved class of estimators of a finite population quantile in sample surveys. *Applied mathematics letters*, *20*(3), 312–315.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(3), 445–458.

Bickel, P. J., & Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The annals of statistics*, (pp. 470–482).

Bogin, B., & Sullivan, T. (1986). Socioeconomic status, sex, age, and ethnicity as determinants of body fat distribution for guatemalan children. *American Journal of Physical Anthropology*, *69*(4), 527–535.

Bohn, M. K., Higgins, V., Kavsak, P., Hoffman, B., & Adeli, K. (2019). High-sensitivity generation 5 cardiac troponin t sex-and age-specific 99th percentiles in the caliper cohort of healthy children and adolescents. *Clinical Chemistry*, *65*(4), 589–591.

Booth, J. G., Butler, R. W., & Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, *89*(428), 1282–1289.

Breidt, F., Opsomer, J., Johnson, A., & Ranalli, M. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology*, *33*(1), 35.

Brewer, K. (2000). Cosmetic calibration with unequal probability sampling. *Survey Methodology*, *25*, 205–212.

Bu, R., Lu, J., Ren, T., Liu, B., Li, X., & Cong, R. (2015). Particulate organic matter affects soil nitrogen mineralization under two crop rotation systems. *PLoS One*, *10*(12), e0143835.

Burtless, G. (1999). Effects of growing wage disparities and changing family composition on the us income distribution. *European Economic Review*, *43*(4-6), 853–865.

Cardot, H., Goga, C., & Shehzad, M. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, *27*(1), 243–260.

Chambers, R., & Clark, R. (2008). Adaptive calibration for prediction of finite population totals. *Survey Methodology*, *34*(2), 163–172.

Chambers, R., Dorfman, A. H., & Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika*, *79*(3), 577–582.

Chambers, R. L., & Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, *73*(3), 597–604.

Chang, T., & Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, *95*(3), 555–571.

Chao, M.-T., & Lo, S.-H. (1994). Maximum likelihood summary and the bootstrap method in structured finite populations. *Statistica Sinica*, (pp. 389–406).

Chauvet, G. (2007). *Méthodes de bootstrap en population finie*. Ph.D. thesis, ENSAI.

Chauvet, G., & Goga, C. (2022). Asymptotic efficiency of the calibration estimator in a highdimensional data setting. *Journal of Statistical Planning and Inference*, *217*, 177–187.

Chauvet, G., et al. (2007). Bootstrap pour un tirage à plusieurs degrés avec échantillonnage à forte entropie à chaque degré. Tech. rep.

Chen, J., & Sitter, R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, (pp. 385–406).

Chen, J., & Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, *12*(4), 1223–1239.

Countouris, N., Jagodzinski, R., Bérastégui, P., De Spiegelaere, S., Degryse, C., Franklin, P., Galgóczi, B., Hoffmann, A., Hernandez, S. L., Müller, T., et al. (2020). Benchmarking working europe 2020.

Crawford, J. R., & Garthwaite, P. H. (2009). Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *The Clinical Neuropsychologist*, *23*(2), 193–204.

Darvas, Z. (2019). Why is it so hard to reach the eu's poverty target? *Social Indicators Research*, *141*(3), 1081–1105.

Decker, R., Haltiwanger, J., Jarmin, R., & Miranda, J. (2014). The role of entrepreneurship in us job creation and economic dynamism. *Journal of Economic Perspectives*, *28*(3), 3–24.

Devaud, D., & Tillé, Y. (2019). Deville and särndal's calibration: revisiting a 25 years old successful optimization problem. *Test*, *28*(4), 1033–1065.

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. In *COMPSTAT*, (pp. 65–76). Springer.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376–382.

Dickens, R., & Manning, A. (2004). Has the national minimum wage reduced uk wage inequality? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *167*(4), 613–626.

Dorfman, A. H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, *35*(1), 29–41.

Dorfman, A. H. (2009). Inference on distribution functions and quantiles. In *Handbook of statistics*, vol. 29, (pp. 371–395). Elsevier.

Dorfman, A. H., & Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, (pp. 1452–1475).

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. annals of statistics, 7, 1-26. *The Annals of Statistics*, *7*(1), 1–26.

Eisenberg, M. E., Neumark-Sztainer, D., Story, M., & Perry, C. (2005). The role of social norms and friends' influences on unhealthy weight-control behaviors among adolescent girls. *Social Science & Medicine*, *60*(6), 1165–1173.

Estevao, V., & Särndal, C. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, *74*(2), 127–147.

Estevao, V. M., & Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, *16*(4), 379.

European Commission (2010a). Europe 2020: A strategy for smart, sustainable and inclusive growth. `https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=LEGISSUM:em0028`. Online; accessed 05 September 2022.

European Commission (2010b). The european platform against poverty and social exclusion: A european framework for social and territorial cohesion. `https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=LEGISSUM:em0046`. Online; accessed 05 September 2022.

Eurostat Experimental statistics (2022). Income inequality and poverty indicators. `https://ec.europa.eu/eurostat/web/experimental-statistics/income-inequality-and-poverty-indicators`. Online; Acceso 18 Septiembre 2022.

Eurostat Products Datasets (2022). Inequality of income distribution s80/s20 income quintile share ratio - eusilc and echp surveys. `https://ec.europa.eu/eurostat/web/products-datasets/-/ilc_pns4`. Online; Acceso 21 Septiembre 2022.

Eurostat Statistics (2020). Living conditions in europe-poverty and social exclusion. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Living_conditions_in_Europe_-_poverty_and_social_exclusion. Online; accessed 29 Septiembre 2022.

Eurostat Statistics (2022). People at risk of poverty or social exclusion. https://ec.europa.eu/eurostat/web/products-datasets/-/tipslc10. Online; accessed 12 Octubre 2022.

Ferri-García, R., & Rueda, M. d. M. (2018). Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *1*, *2*(1), 159–162.

Gelman, A., Kenworthy, L., & Su, Y.-S. (2010). Income inequality and partisan voting in the united states. *Social Science Quarterly*, (pp. 1203–1219).

Gross, S. (1980). Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, vol. 1814184. American Statistical Association Alexandria, VA.

Guggemos, F., & Tillé, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference*, *140*(11), 3199–3212.

Guio, A.-C., Marlier, É., & Nolan, B. (2021). *Improving the understanding of poverty and social exclusion in Europe*. Publications Office of the European Union Luxembourg.

Harms, T., & Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, *32*(1), 37–52.

INE (2015). Continuous register statistics. https://www.ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736177012&menu=resultados&secc=1254736195462&idp=1254734710990#!tabs-1254736195462*https* : *//www.ine.es/dyngs/INEbase/en/operacion.htm?c* = *Estadistica_C&cid* = 1254736177012&*menu* = *resultados*. Online; accessed 01 Noviembre 2022.

INE (2022). Encuesta de condiciones de vida. https://www.ine.es/daco/daco42/condivi/ecv_metodo.pdf. Online; accessed 01 Noviembre 2022.

Jo, B., Laitila, T., et al. (2015). Comparisons of some weighting methods for non-response adjustment. *Lithuanian Journal of Statistics*, *54*(1), 69–83.

Jones, A. F., & Weinberg, D. H. (2000). *The changing shape of the nation's income distribution, 1947-1998*. 204. US Department of Commerce, Economics and Statistics Administration, US ....

Kim, J. K., & Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, *78*(1), 21–39.

Kimbro, R. T., Brooks-Gunn, J., & McLanahan, S. (2011). Young children in urban areas: links among neighborhood characteristics, weight status, outdoor play, and television watching. *Social Science & Medicine*, *72*(5), 668–676.

Kott, P. S., & Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. In *Survey Research Methods*, vol. 6, (pp. 105–111).

Kott, P. S., & Liao, D. (2015). One step or two? calibration weighting from a complete list frame with nonresponse. *Survey Methodology*, *41*(1), 165–182.

Kott, P. S., & Liao, D. (2017). Calibration weighting for nonresponse that is not missing at random: Allowing more calibration than response-model variables. *Journal of Survey Statistics and Methodology*, *5*(2), 159–174.

Kovacevic, M. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. vol. 47, (pp. 139–144).

Kuk, A. Y., & Mak, T. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society: Series B (Methodological)*, *51*(2), 261–269.

Lafferty, P., & McCormack, K. (2015). A review of the sampling and calibration methodology of the survey on income and living conditions (silc) 2010-2013. `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f50fa2b2892cbfd9698e32dcc7099cf5faef0dde`.

Le Guennec, J., & Sautory, O. (2002). Calmar 2: Une nouvelle version de la macro calmar de redressement d'échantillon par calage. *Journées de Méthodologie Statistique, Paris. INSEE*.

Lesage, É., Haziza, D., & D'Haultfœuille, X. (2019). A cautionary tale on instrumental calibration for the treatment of nonignorable unit nonresponse in surveys. *Journal of the American Statistical Association*, *114*(526), 906–915.

Lundström, S., & Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of official statistics*, *15*(2), 305.

Machin, S., Manning, A., & Rahman, L. (2003). Where the minimum wage bites hard: Introduction of minimum wages to a low wage sector. *Journal of the European Economic Association*, *1*(1), 154–180.

Martínez, S., Illescas, M., Martínez, H., & Arcos, A. (2020). Calibration estimator for head count index. *International Journal of Computer Mathematics*, *97*(1-2), 51–62.

Martínez, S., Rueda, M., Arcos, A., & Martínez, H. (2010). Optimum calibration points estimating distribution functions. *Journal of computational and applied mathematics*, *233*(9), 2265–2277.

Martínez, S., Rueda, M., Arcos, A., Martínez, H., & Muñoz, J. (2012). On determining the calibration equations to construct model-calibration estimators of the distribution function. *Revista matemática complutense*, *25*(1), 87–95.

Martínez, S., Rueda, M., Arcos, A., Martínez, H., & Sánchez-Borrego, I. (2011). Post-stratified calibration method for estimating quantiles. *Computational statistics & data analysis*, *55*(1), 838–851.

Martínez, S., Rueda, M., & Illescas, M. (2022). The optimization problem of quantile and poverty measures estimation based on calibration. *Journal of Computational and Applied Mathematics*, *405*, 113054.

Martínez, S., Rueda, M., Martínez, H., & Arcos, A. (2015). Determining p optimum calibration points to construct calibration estimators of the distribution function. *Journal of computational and applied mathematics*, *275*, 281–293.

Martínez, S., Rueda, M., Martínez, H., & Arcos, A. (2017). Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function. *Journal of Computational and Applied Mathematics*, *318*, 444–459.

Martínez Puertas, S., & Martínez Puertas, H. (2013). Aplicación de técnicas de calibración en la estimación de líneas de pobreza. *Estadística española*, *55*(182), 323–336.

Mayor-Gallego, J., Moreno-Rebollo, J., & Jiménez-Gamero, M. (2019). Estimation of the finite population distribution function using a global penalized calibration method. *AStA Advances in Statistical Analysis*, *103*(1), 1–35.

McConville, K., Breidt, F., Lee, T., & Moisen, G. (2017). Model assisted survey regression estimation with the lasso. *J. Surv. Stat. Methodol*, *5*(2), 131–158.

Meglio, E. D. (Ed.) (2018). *Living conditions in Europe — 2018 edition*. Publications Office of the European Union, Luxembourg.

Memobust (2014). Weighting and estimation - calibration. ://ec.europa.eu/eurostat/cros/content/calibration-method$_en. Online; accessed 11 Noviembre 2022$.

Metcalf, D. (2008). Why has the british national minimum wage had little or no impact on employment? *Journal of Industrial Relations*, *50*(3), 489–512.

Meyer, B. D., & Sullivan, J. X. (2012). Identifying the disadvantaged: Official poverty, consumption poverty, and the new supplemental poverty measure. *Journal of Economic Perspectives*, *26*(3), 111–36.

Montanari, G. E., & Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, *100*(472), 1429–1442.

Morales, D., Rueda, M. d. M., & Esteban, D. (2018). Model-assisted estimation of small area poverty measures: an application within the valencia region in spain. *Social Indicators Research*, *138*(3), 873–900.

Mukhopadhyay, P. (2012). *Topics in survey sampling*. Springer New York, NY.

Muñoz, J., Álvarez-Verdejo, E., García-Fernández, R., & Barroso, L. (2015). Efficient estimation of the headcount index. *Social Indicators Research*, *123*(3), 713–732.

Nascimento Silva, P., & Skinner, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, *23*(1), 23–32.

Nickell, S. (2004). Poverty and worklessness in britain. *The Economic Journal*, *114*(494), C1–C25.

Ranalli, M. G., Arcos, A., Rueda, M. d. M., & Teodoro, A. (2016). Calibration estimation in dual-frame surveys. *Statistical Methods & Applications*, *25*(3), 321–349.

Rao, J., Kovar, J., & Mantel, H. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, (pp. 365–375).

Rodrigues, J., Bolfarine, H., & Rogatko, A. (1985). A general theory of prediction in finite populations. *International Statistical Review/Revue Internationale de Statistique*, (pp. 239–254).

Rota, B. (2017). Variance estimation in two-step calibration for nonresponse adjustment. *South African Statistical Journal*, *51*(2), 361–374.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, *57*(2), 377–387.

Rueda, M. (2019). Comments on: Deville and särndal's calibration: revisiting a 25 years old successful optimization problem. *Test*, *28*(4), 1077–1081.

Rueda, M., Martinez, S., Arcos, A., & Munoz, J. (2009). Mean estimation under successive sampling with calibration estimators. *Communications in Statistics—Theory and Methods*, *38*(6), 808–827.

Rueda, M., Martínez, S., & Illescas, M. (2021). Treating nonresponse in the estimation of the distribution function. *Mathematics and Computers in Simulation*, *186*, 136–144.

Rueda, M., Martinez, S., Martinez, H., & Arcos, A. (2006). Mean estimation with calibration techniques in presence of missing data. *Computational Statistics & Data Analysis*, *50*(11), 3263–3277.

Rueda, M., Martínez, S., Martínez, H., & Arcos, A. (2007a). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, *137*(2), 435–448.

Rueda, M., Martínez-Puertas, S., Martínez-Puertas, H., & Arcos, A. (2007b). Calibration methods for estimating quantiles. *Metrika*, *66*(3), 355–371.

Särndal, C. E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, *67*(3), 639–650.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey methodology*, *33*(2), 99–119.

Särndal, C.-E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.

Sedransk, N., & Sedransk, J. (1979). Distinguishing among distributions using data from complex sample designs. *J Am Stat Assoc*, *74*(368), 754–760.

Shrider, E. A., Kollar, M., Chen, F., Semega, J., et al. (2021). Income and poverty in the united states: 2020. *US Census Bureau, Current Population Reports*, (P60-273).

Silva, P., & Skinner, C. J. (1995). Estimating distribution functions with auxiliary information using poststratification. *Journal of Official Statistics*, *11*(3), 277–294.

Singh, H. P., Singh, S., & Kozak, M. (2008). A family of estimators of finite-population distribution function using auxiliary information. *Acta applicandae mathematicae*, *104*(2), 115–130.

Singh, S. (2001). Generalized calibration approach for estimating variance in survey sampling. *Annals of the Institute of Statistical Mathematics*, *53*(2), 404–417.

Sompolska-Rzechu la, A., & Kurdyś-Kujawska, A. (2022). Assessment of the development of poverty in eu countries. *International Journal of Environmental Research and Public Health*, *19*(7), 3950.

Tellez-Plaza, M., Navas-Acien, A., Crainiceanu, C. M., & Guallar, E. (2008). Cadmium exposure and hypertension in the 1999–2004 national health and nutrition examination survey (nhanes). *Environmental Health Perspectives*, *116*(1), 51–56.

Tillé, Y., & Matei, A. (2021). *sampling: Survey Sampling*. R package version 2.9. URL `https://CRAN.R-project.org/package=sampling`

Vander Wal, J. S., & Mitchell, E. R. (2011). Psychological complications of pediatric obesity. *Pediatric Clinics*, *58*(6), 1393–1401.

Vanderhoeft, C. (2001). *Generalised calibration at Statistics Belgium: SPSS module g-CALIB-S and current practices*. Bruxelles : Inst. National de Statistique.

Wang, S., & Dorfman, A. H. (1996). A new estimator for the finite population distribution function. *Biometrika*, *83*(3), 639–652.

Wilson, R., Fleming, Z. L., Monks, P., Clain, G., Henne, S., Konovalov, I., Szopa, S., & Menut, L. (2012). Have primary emission reduction measures reduced ozone across europe? an analysis of european rural background ozone trends 1996–2005. *Atmospheric Chemistry and Physics*, *12*(1), 437–454.

Wolford, S., Schroer, R., Gohs, F., Gallo, P., Brodeck, M., Falk, H., & Ruhren, R. (1986). Reference range data base for serum chemistry and hematology values in laboratory animals. *Journal of Toxicology and Environmental Health, Part A Current Issues*, *18*(2), 161–188.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, *90*(4), 937–951.

Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, *96*(453), 185–193.

Parte II

Apéndices

Apéndice A1

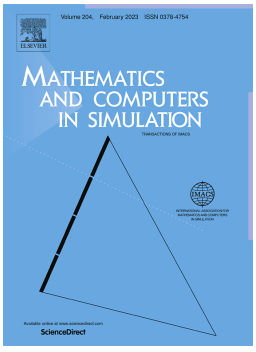
Treating nonresponse in the estimation of the distribution function

Rueda, María del Mar; Martínez, Sergio; Illescas, María Dolores (2021)

Treating nonresponse in the estimation of the distribution function.

Mathematics and Computers in Simulation, Vol. 186, pp. 136–144.

DOI: 10.1016/j.matcom.2020.07.027



MATHEMATICS, APPLIED			
JCR Year	Impact factor	Rank	Quartile
2021	3.601	18/267	Q1

Abstract

The estimation of a finite population distribution function is considered when there are missing data. Calibration adjustment is used for dealing with nonresponse at the estimation stage. Several procedures are proposed and compared. A numerical study is carried out to evaluate the performances of estimators. Computational problems with the implementation of the proposed calibration estimators are also considered.

1. Introduction

Surveys are a common method of data collection in economics and social sciences, but they often suffer from the problem of nonresponse which can produce biases in estimations and an increase in sampling variance if missing data follows any pattern. The standard statistical procedures developed for data with no missing values cannot be immediately and straightforwardly applied for deducing inferences in this situation.

Weighting is widely applied in surveys to adjust for nonresponse. Many different proposals for nonresponse weighting have been considered (see.eg. Beaumont (2005),Chang & Kott (2008),Kott & Liao (2012),Kott & Liao (2015), Kott & Liao (2017), Lundström & Särndal (1999), Jo et al. (2015), Lesage et al. (2019)) in the estimation of linear parameters as total or mean, but lesser effort has been devoted in the development of efficient methods for estimating a population distribution. The distribution function is a relevant tool when the variable of interest is a measure of wages or income, since it is needed to calculate many poverty measures (the poverty line, the low income proportion, the poverty gap ...) For these reasons, estimation of the distribution function is an important issue in sample surveys that has received much attention in the last years. On the contrary, to best of our knowledge, its estimation in the presence of missing data is a issue less investigated in the previous research. Whereas a extensive literature is available on estimation of population mean under non-response, lesser effort has been devoted in the development of efficient methods for the estimation of population distribution function.

The purpose of this paper is to estimate the distribution function in presence of missing data using the calibration method under a general sampling design. To the best of our knowledge, this is the first time that calibration techniques have been employed to remove the bias of non response in the estimation of the distribution function.

2. Estimating the distribution function when there are missing values

Given a finite population $U = \{1, \ldots, N\}$ with N different units and a sampling design d defined in U with first-order inclusion probability $\pi_i \geq 0$ and $d_i = \pi_i^{-1}$ the sampling design-basic weight for unit $i \in U$. In the presence of unit nonresponse, the character under study, say y, is observed for a subset of the original sample s. Thus, if we assume missing data on the sample s, it can be divided into the disjoint sets:

$$s_r = \{i \in s/\text{i responds}\} \text{ and } s_m = \{i \in s/\text{i does not respond}\},$$

with s_r, the respondent sample is of size r, and s_m is of size $n - r$.

Let y_i be the value of the character under study, say y, for the ith population unit. Our aim is to estimate the finite population distribution function (f.d.f.) of the study variable y, given by

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k), \text{ with } \Delta(t - y_k) = \begin{cases} 0 & \text{if } t < y_k \\ 1 & \text{if } t \geq y_k \end{cases}$$

The design-unbiased Horvitz-Thompson estimator of $F_y(t)$ defined by

$$\widehat{F}_Y(t) = \frac{1}{N} \sum_{k \in s} d_k \Delta(t - y_k).$$

is impossible to compute in the presence of unit nonresponse, and a naive estimator of $F_y(t)$ is:

$$\widehat{F}_{YH}(t) = \frac{1}{N_r} \sum_{k \in s_r} d_k \Delta(t - y_k)$$

where N_r is the size of the population that would have responded if sampled, a quantity that is very rarely known in practice. When N_r is not known, one can use the Hajek estimator

$$\widehat{F}_{YHa}(t) = \frac{\sum_{k \in s_r} d_k \Delta(t - y_k)}{\sum_{k \in s_r} d_k}.$$

These estimators may lead to biased estimates because certain specific groups can be substantially under-represented. These errors can be overcome by the use of reweighting. When weighting is used, a set of weights is determined with the aid of the available auxiliary information, and estimation is carried out by applying the weights to the y-values for the responding elements. Calibration adjustment, initially conceived for correcting sampling errors (Deville & Särndal (1992)), is currently one of the most appealing techniques for nonresponse adjustment (Särndal & Lundström (2005)).

We will use a twofold process: the sample s is first selected from the population U, then the response set s_r is realized as a subset of s. We assume that elements respond independently and the response distribution has first-order response probabilities $P(k \in s_r/s) = p_k$, positives. These response probabilities (whose true values are unknown) are estimated by $\hat{p}_k$ and we can obtain a two-phase nonresponse adjusted estimator of the distribution function by replacing the original design weights by $d_k^o = (\pi_k \hat{p}_k)^{-1}$, that is:

$$\widehat{F}_{Yw}(t) = \sum_{k \in s_r} d_k^o \Delta(t - y_k).$$

For it, we assume the existence of auxiliary information relative to several variables related to the main variable y, $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kJ})'$. The values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ are known for the entire population but y_k is known only if the kth unit is selected in the sample s_r.

3. Calibration weighting for the estimation of the distribution function with unit nonresponse.

Harms & Duchesne (2006) and Rueda et al. (2007a) use different ways to implement the calibration approach in the estimation of the distribution function and the quantiles. The computationally simpler method proposed by Rueda et al. (2007a) uses the calibration with respect to the predicted values of the variable of interest y. We use this methodology and we define a pseudo-variable $\tilde{y}_k = \widehat{\beta}^T \mathbf{x}_k$ for $k = 1, 2, ...N$, where $\widehat{\beta} = \left(\sum_{j \in s_r} d_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \cdot \sum_{j \in s_r} d_j \mathbf{x}_j y_j$.

Given a distance measure $G(w_k, d_k)$, the calibration process consists in finding the solution of the following minimization problem

$$\min_{w_k} \sum_{s_r} G(w_k, d_k) \tag{1.1}$$

while respecting the calibration equation

$$\frac{1}{N} \sum_{k \in s_r} w_k \Delta(\mathbf{t} - \tilde{y}_k) = F_{\tilde{y}}(\mathbf{t}) \tag{1.2}$$

where $\mathbf{t} = (t_1, t_2, \ldots, t_P{}')$, the term $F_{\tilde{y}}(\mathbf{t})$ denotes the f.d.f. of the pseudo-variable $\tilde{y}$ evaluated at $\mathbf{t}$ (see Martínez et al. (2010)) and t_j for $j = 1, 2, \ldots, P$ are points that we choose arbitrarily and assume that $t_1 < t_2 < \ldots < t_P$.

We can define the calibration estimator

$$\widehat{F}_{cal}(t) = \frac{1}{N}\sum_{k\in s_r} w_k \Delta(t - y_k) \tag{1.3}$$

What form should the weight take? It should reflect the known individual characteristics of the element $k \in r$, summarized by the vector value $\Delta(\mathbf{t} - \tilde{y}_k)$. A common way to compute calibration weights is linearly (using the chi-square distance, Lundström & Särndal (1999)) that produces the weights $w_k = d_k(1 + \lambda^T \Delta(\mathbf{t} - \tilde{y}_k))$. The weights w_k implicitly estimates the inverse response probability $1/p_k$, thus we acts as if $\pi_k p_k$ is the true selection probability of element k. Consequently this calibration approach has assumed a nonresponse model $p_k = 1/(1+\lambda^T \Delta(\mathbf{t} - \beta^T \mathbf{x}_k))$. This model is difficult to deal with since the function is not differentiable on $\mathbf{x}_k$ and assume that the response mechanism depend on all auxiliary variables.

Now we consider a more flexible model and we propose a calibration method that allows the variables modeling the response mechanism to be different from the benchmark variables in the calibration equation. Thus we define this two-step calibration method:

Step 1: Adjusting the bias of nonresponse by linear calibration.

Consider the M vector of explanatory model variables, $\mathbf{x}_k^*$ which population totals $\sum_{\mathcal{U}} \mathbf{x}_k^*$ are know. The calibration under the restrictions $\sum_{s_r} w_k^{(1)} \mathbf{x}_k^* = \sum_{\mathcal{U}} \mathbf{x}_k^*$ yields the calibrations weights $w_k^{(1)} = g_k^{(1)} * d_k, k = 1, \ldots, s_r$. Then, each unit in the sample has a weight that corrects the bias of lack of response.

Step 2: Adjusting the sample weights for the estimation of the f.d.f.

The auxiliary information of the calibration variables $\mathbf{x}$ is incorporated through the calibrated weights $w_k^{(2)} = g_k^{(2)} * w_k^{(1)}$ obtained with the restrictions $\sum_{s_r} w_k^{(2)} \Delta(\mathbf{t} - \tilde{y}_k) = F_{\tilde{y}}(\mathbf{t})$.

The two step calibration estimator proposed is

$$\hat{F}_{calTS}(t) = \frac{1}{N}\sum_{s_r} w_k^{(2)} \Delta(t - y_k) = \sum_{s_r} g_k^{(2)} g_k^{(1)} * d_k \Delta(t - y_k) \tag{1.4}$$

Note that the vector of model variables $\mathbf{x}_k^*$ and the vector of calibration variables $\mathbf{x}$ can be the same, can have some common component or be completely different

4. Calibration with model and calibration variables

In this section we consider a calibration approach similar to that used by Kott & Liao (2015) and Kott & Liao (2017) for the estimation of the population total. We also allow the variables modelling to be different from the calibration variables.

The probability of nonresponse can be modeled by $p_k = f(\gamma^T \mathbf{x}_k^*)$ for some vector parameter γ, where

$h(\cdot) = 1/f(\cdot)$ is a known and everywhere monotonic and twice differentiable function and the vector $\mathbf{x}_k^*$ is the vector with the model variables. Some examples of usual models for the probability of response are: $p_k = \frac{1+exp(\gamma^T\mathbf{x}_k)^*/u}{l+exp(\gamma^T\mathbf{x}_k)^*}$ (the logit (l,u) method), $p_k = \frac{1}{exp(-\gamma^T\mathbf{x}_k^*)}$ (the raking model) and $p_k = \frac{exp(\gamma^T\mathbf{x}_k^*)}{1+exp(\gamma^T\mathbf{x}_k^*)}$ (the logistic-response model, a special case of logit (l,u) method, where $u = \infty$, $c = 2$ and $l = 1$ (Kott & Liao (2012))). The response probability is assumed independent of the survey variable of interest, which is known as missing at random (MAR) assumption.

We denote as $\mathbf{z}_k = \Delta(\mathbf{t} - \tilde{y}_k)$. We will use the vector $\mathbf{z}_k$ in the benchmark. Now, we generate calibrated weights by imposing the functional form $w_k = \frac{d_k}{f(\hat{\gamma}^T\mathbf{x}_k^*)}\mathbf{z}_k$. The calibration equation is given by:

$$\frac{1}{N}\sum_{s_r}\frac{d_k}{f(\hat{\gamma}^T\mathbf{x}_k^*)}\mathbf{z}_k = \frac{1}{N}\sum_{s_r} d_k h(\hat{\gamma}^T\mathbf{x}_k^*)\mathbf{z}_k = F_{\tilde{y}}(\mathbf{t}) \tag{1.5}$$

where $\hat{\gamma}$ is a consistent estimator of vector γ and the resulting calibration estimator is given by

$$\hat{F}_{calI}(t) = \frac{1}{N}\sum_{s_r} d_k\frac{1}{\hat{p}_k}\Delta(t - y_k) = \frac{1}{N}\sum_{s_r} d_k h(\hat{\gamma}^T\mathbf{x}_k^*)\Delta(t - y_k).$$

Some considerations about the number of variables in the non-response model M and the number of calibration restrictions P should be taken into account to obtain the estimator $\hat{F}_{calI}(t)$. In Deville (2000), the number of model and calibration variables needed to be the same, that is $M = P$. This issue may limit the number of calibration restrictions in practice. We denote by $\hat{F}_{calID}$ the estimator $\hat{F}_{calI}$ when we use the same number of model and calibration variables.

Kott & Liao (2017) extended the Deville's weighting approach to the case where there are more calibration variables than model variables ($P > M$)through two alternatives. For it, their first extension does not look for an estimation $\widehat{\gamma}$ that satisfies the calibration equation (2.5), but it looks for an estimation $\widehat{\gamma}$ that minimizes $\mathbf{v}^T Q\mathbf{v}$ for some symmetric and positive definite matrix Q with dimension P, where

$$\mathbf{v} = \frac{1}{N}\left(\sum_{s_r} d_k h(\hat{\gamma}^T\mathbf{x}_k^*)\mathbf{z}_k - \sum_{s} d_k\tilde{\mathbf{z}}_k\right)$$

This minimization problem implies the reformulated calibration equation:

$$\frac{1}{N}\sum_{s_r} w_k^{(3)}\tilde{\mathbf{z}}_k = \frac{1}{N}\sum_{s_r} d_k h(\hat{\gamma}^T\mathbf{x}_k^*)\tilde{\mathbf{z}}_k = \frac{1}{N}\sum_{s} d_k\tilde{\mathbf{z}}_k \tag{1.6}$$

where $\tilde{\mathbf{z}}_k = A\cdot\mathbf{z}_k$ with $A = \left[\frac{1}{N}\sum_{s_r} d_i h'(\hat{\gamma}^T\mathbf{x}_i^*)x_i^*\mathbf{z}_i^T Q\right]$.

Thus, the first alternative in Kott & Liao (2017) considers iterative process to find an estimation $\hat{\gamma}$ and

Q that satisfied equation (1.6) such that $Q = H^{-1}$ with

$$H = \frac{1}{N}\sum_{s_r} d_k h^{'}(\hat{\gamma}^T \mathbf{x}_k^*)\mathbf{z}_k(\mathbf{z}_k)^T \tag{1.7}$$

In this alternative, each variable included in the vector $\tilde{\mathbf{z}}_k$ is a prediction for the corresponding variable in $\mathbf{x}_k^*$.

The second alternative proposed in Kott & Liao (2017) is a variant of the component reduction technique based on equation (1.6) with $\tilde{\mathbf{z}}_k = A_0 \cdot \mathbf{z}_k$ where

$$A_0^T = \left(\sum_{s_r} \mathbf{z}_j(\mathbf{z}_j)^T\right)^{-1} \sum_{s_r} \mathbf{z}_j(\mathbf{x}_j^*)^T$$

Consequently, this alternative does not require iteration or even rely on finding a matrix Q.

The estimator $\hat{F}_{calI}$ based on the first alternative of Kott and Liao approach (Kott & Liao (2017)) is denoted by $\hat{F}_{calIKL1}$ and the estimator $\hat{F}_{calI}$ based on second alternative, is denoted by $\hat{F}_{calIKL2}$.

5. Properties of the calibrated estimators of the distribution function.

For an estimator $\hat{F}(t)$ of $F(t)$ to be a genuine distribution function it should verify:

i. $\widehat{F}_{cal}(t)$ is continuous on the right,

ii. $\widehat{F}_{cal}(t)$ is monotone nondecreasing,

iii. a) $\lim\limits_{t\to-\infty} \widehat{F}_{cal}(t) = 0$ and b) $\lim\limits_{t\to+\infty} \widehat{F}_{cal}(t) = 1$.

It's easy to verify that conditions *i)* and *iii.a)* are satisfied for all the proposed estimators. $\hat{F}_{calTS}(t)$ meet the condition iii. b) if we take t_P such that $F_{\tilde{y}}(t_P) = 1$ (see Rueda et al. (2007a)) but in general this estimator is not monotone nondecreasing. With respect to the calibration estimators based on model and calibration variables $\widehat{F}_{calID}$, $\widehat{F}_{calIKL1}$ and $\widehat{F}_{calIKL2}$, the calibration weights $\omega_k \geq 0$ for logit, raking and logistic methods and therefore, the estimators are nondecreasing. Like the previous case, for $\widehat{F}_{calID}$, if we take t_P such that $F_{\tilde{y}}(t_P) = 1$, then $\lim_{t\to+\infty} \widehat{F}_{calID}(t) = 1$. On the other hand, for the estimators $\widehat{F}_{calIKL1}$ and $\widehat{F}_{calIKL2}$, Theorem 1 establish the conditions to satisfy $\lim_{t\to+\infty} \widehat{F}_y(t) = 1$.

Theorem 1: If a component of the vector x_k^* contains all 1's and t_P should be sufficiently large, the estimators $\widehat{F}_{calIKL1}$ and $\widehat{F}_{calIKL2}$ satisfy the condition iii. b).

Proof:

We denote $\Delta(\mathbf{t}-\tilde{y}_k)^T = \Big(\Delta(t_1-\tilde{y}_k),\ldots,\Delta(t_{P-1}-\tilde{y}_k),\Delta(t_P-\tilde{y}_k)\Big) = (\Delta_{P-1}^T,1)$ and $(x_k^*)^T = (1,x_{2k},\ldots,x_{Mk}) = \Big(1,(x_k^\circ)^T\Big)$.

For the estimator $\widehat{F}_{calIKL1}(t)$, the calibration weights ω_k satisfies (1.6) with Q given by (1.7). It is clear that the matrix Q and Q^{-1} can be expressed by

$$Q=\begin{pmatrix} Q_{(P-1)\times(P-1)} & D_{(P-1)\times 1} \\ T_{1\times(P-1)} & C_{1\times 1}\end{pmatrix} \quad ; \quad Q^{-1}=\begin{pmatrix} \Gamma_{(P-1)\times(P-1)} & \Psi_{(P-1)\times 1} \\ (\Psi_{(P-1)\times 1})^T & \Upsilon_{1\times 1}\end{pmatrix}$$

with $\Psi_{(P-1)\times 1}=\frac{1}{N}\sum_{s_r} d_k h^{'}(\hat{\gamma}^T\mathbf{x}_k^*)\Delta_{P-1}$; $\Upsilon_{1\times 1}=\frac{1}{N}\sum_{s_r} d_k h^{'}(\hat{\gamma}^T\mathbf{x}_k^*)$ and $\Gamma_{(P-1)\times(P-1)}$ is given by equation (1.7) based on Δ_{P-1}. From $Q^{-1}\cdot Q = I_{P\times P}$, we have

$$(\Psi_{(P-1)\times 1})^T Q_{(P-1)\times(P-1)} + \Upsilon_{1\times 1} T_{1\times(P-1)} = 0_{1\times(P-1)} \tag{1.8}$$

$$(\Psi_{(P-1)\times 1})^T D_{(P-1)\times 1} + \Upsilon_{1\times 1} C_{1\times 1} = 1. \tag{1.9}$$

From (1.8) and (1.9),we have:

$$A=\left(\frac{1}{N}\sum_{s_r} d_i h^{'}(\hat{\gamma}^T\mathbf{x}_i^*)x_i^*\Delta(\mathbf{t}-\tilde{y}_k)^T Q\right)=\begin{pmatrix} 0_{1\times(P-1)} & 1 \\ A_{(M-1)\times(P-1)} & B_{(M-1)\times 1}\end{pmatrix}$$

with $B_{(M-1)\times 1} = (\Phi_{(P-1)\times(M-1)})^T D_{(P-1)\times 1} + \chi_{(M-1)\times 1} C_{1\times 1}$

$$A_{(M-1)\times(P-1)} = (\Phi_{(P-1)\times(M-1)})^T Q_{(P-1)\times(P-1)} + \chi_{(M-1)\times 1} T_{1\times(P-1)}$$

$$(\Phi_{(P-1)\times(M-1)})^T = \frac{1}{N}\sum_{s_r} d_i h^{'}(\hat{\gamma}^T\mathbf{x}_i^*)x_i^\circ\Delta_{P-1}^T \quad ; \quad \chi_{(M-1)\times 1} = \frac{1}{N}\sum_{s_r} d_i h^{'}(\hat{\gamma}^T\mathbf{x}_i^*)x_i^\circ$$

Consequently, $\tilde{\mathbf{z}}_k$ is given by $\tilde{\mathbf{z}}_k = A\Delta(\mathbf{t}-\tilde{y}_k) = \begin{pmatrix} 1 \\ Z_{(M-1)\times 1}\end{pmatrix}$ with

$$Z_{(M-1)\times 1} = A_{(M-1)\times(P-1)}\Delta_{P-1} + B_{(M-1)\times 1}$$

and the property iii. b) is fullfilled.

For the $\widehat{F}_{calIKL2}$ estimator, matrix A_0 can be expressed by $A_0 = \begin{pmatrix} A_{01} \\ A_{0(M-1)} \end{pmatrix}$ where

$$A_{01} = \left(\sum_{s_r} \Delta(\mathbf{t} - \tilde{y}_k)^T \right) \left(\sum_{s_r} \Delta(\mathbf{t} - \tilde{y}_k) \Delta(\mathbf{t} - \tilde{y}_k)^T \right)^{-1}$$

$$A_{0(M-1)} = \left(\sum_{s_r} x_k^0 \Delta(\mathbf{t} - \tilde{y}_k)^T \right) \left(\sum_{s_r} \Delta(\mathbf{t} - \tilde{y}_k) \Delta(\mathbf{t} - \tilde{y}_k)^T \right)^{-1}$$

For t_P sufficiently large, $\Delta(t_P - \tilde{y}_k) = 1$ for all $k \in U$ and we have

$$A_{01} = \left(\sum_{s_r} \Delta(t_P - \tilde{y}_k) \Delta(\mathbf{t} - \tilde{y}_k)^T \right) \left(\sum_{s_r} \Delta(\mathbf{t} - \tilde{y}_k) \Delta(\mathbf{t} - \tilde{y}_k)^T \right)^{-1} = \begin{pmatrix} 0_{1\times(P-1)} & 1 \end{pmatrix}$$

Therefore, the vector $\tilde{z}_k$ is given by $\tilde{z}_k = A_0 \cdot \Delta(\mathbf{t} - \tilde{y}_k) = \begin{pmatrix} 1 \\ A_{0(M-1)} \cdot \Delta(\mathbf{t} - \tilde{y}_k) \end{pmatrix}$ and then $\widehat{F}_{calIKL2}(t)$ satisfies property iii. b).

The property ii) could be achieved by the procedure described in Rao et al. (1990). This procedure, for a general estimator $\widehat{F}_y$, is defined in the following way:

$$\tilde{F}_y(y_{[1]}) = \widehat{F}_y(y_{[1]}), \quad \tilde{F}_y(y_{[i]}) = \text{máx}\{\widehat{F}_y(y_{[i]}), \tilde{F}_y(y_{[i-1]})\} \quad i = 2, \ldots, r. \tag{1.10}$$

On the other hand, (Lesage et al. (2019)) established sufficient conditions for consistency of estimators based on model and calibration variables. Thus, under these conditions, the estimators $\widehat{F}_{calID}$, $\widehat{F}_{calIKL1}$ and $\widehat{F}_{calIKL2}$ meet consistency.

Regarding this issue, the conditions required to fulfill the properties of distribution function are not contradictory with respect to the conditions given in (Lesage et al. (2019)) and they are less restrictive than conditions from (Lesage et al. (2019)). In fact, the condition iv) of Assumption 4 from (Lesage et al. (2019)) requires $\mathbb{E}(|z_{j,k}|) < \infty$ for all components of z_k. If t_P is sufficiently large, the last component of z_k meets the condition iv). Also, under the conditions established by Theorem 1, the first component of $\tilde{z}_k$ (for $\widehat{F}_{calIKL1}$ and for $\widehat{F}_{calIKL2}$) meets the condition iv). The same occurs for the condition ii) of Assumption 5 from (Lesage et al. (2019)).

6. Simulation study

We have performed a Monte Carlo simulation study where we compare the precision of the proposed estimators with others estimators of the distribution function. All the estimators included are programmed by routines in R.

6.1. Some computational aspects

The calibration estimator $\hat{F}_{calTS}$ is programmed with routines based on the "calib" function of the package "sampling" (Tillé & Matei (2021)). The "gencalib" function of the package sampling is used for obtain the calibrated weights in $\hat{F}_{calID}$. The raking and logit (l, u) methods are available in the "gencalib"function and we have obtained two versions of this estimator based on the available methods. For the estimator $\hat{F}_{calIK1}$, we have programmed a routine that develops the Newthon-Rhapson method described in Chang & Kott (2008) and we have also obtained two versions (raking and logit (l, u)). With respect to the estimator $\hat{F}_{calIK2}$, we have also obtained two versions and for this we only needed to program a routine for the reduction of components and directly apply the original function "gencalib". For all versions of estimators based on logit (l, u) method, we used the following parameters $u = 10$; $l = 0$ and $c = 1$. Initially, a version of the estimators for the logistic-response model was considered in the simulation study but we finally decided to exclude it because of in many cases, this method did not converge.

6.2. Data

A fictitious population was simulated. The population size is $N = 5000$ and six variables were included in the study: age, nationality (native/non-native), gender, weight, access to the Internet (yes/no). These variables are generated to make its similar to the Spanish population pyramid. The study variable y is given by $y_k = 3 + 5 \cdot Internet + Age/5 + \varepsilon_k$ where ε_k are independent identically distributed random variables with distribution $\varepsilon_k \sim N(0, 0{,}1)$.

First we considered a raking non-response mechanism given by $p_k = \frac{1}{exp(-0{,}1-Internet)}$. Thus, we consider $(x_k^*)^T = (1, Internet)$ and the vector of calibration variables is $(z_k)^T = (1, Internet, Weight)$. Next, we consider a logistic non-response model based on the variable "Age": $p_k = \frac{exp(-3+0{,}1\cdot Age)}{1+exp(-3+0{,}1\cdot Age)}$. In this case, $(z_k)^T = (1, Age, Weight)$.

It is important to note that in both cases, the target variable is not directly related to the non-response mechanism, but this mechanism can be explained by some auxiliary variables configuring a Missing At Random (MAR) situation. We have not considered mechanism Missing Completely At Random (MCAR) since in these situations the estimators are asymptotically unbiased.

6.3. Results

The estimators considered are the Horvitz-Thompson estimator $\widehat{F}_{HT}(t)$, the Chamber-Dunstan estimator $\widehat{F}_{CD}(t)$ (Chambers & Dunstan (1986)), the ratio estimator $\widehat{F}_r(t)$ and the Rao, Kovar and Mantel estimator $\widehat{F}_{RKM}(t)$ (Rao et al. (1990)) based on the respondent sample s_r. The Chamber-Dunstan estimator $\widehat{F}_{CD}(t)$ and the Rao, Kovar and Mantel estimator $\widehat{F}_{RKM}(t)$ are model-based estimator based on the following superpopulation model:

$$y_k = \kappa x_k + \mathrm{v}(x_k)u_k \quad k = 1, \ldots, N \tag{1.11}$$

where κ is an unknown parameter, v is a known, strictly positive function and u_k are independent and identically distributed random variables with zero mean. See (Chambers & Dunstan (1986)) and (Rao et al. (1990)) respectively for further details. In the simulation study, we considered in the superpopulation model (1.11) that $\mathrm{v}(x_k) = 1$ for all unit k.

We drawn 10000 samples with several sizes by simple random sampling without replacement (see Table 1), both for the raking and for the logistic non-response mechanism. For each sample and for each estimator, estimates of the distribution function $F(t)$ were calculated for 11 different values of t, namely all deciles and quartiles $Q_y(0{,}25)$ and $Q_y(0{,}75)$. The performance of all the estimators is measured by means of the average relative bias (AVRB) and the average relative efficiency (AVRE), given respectively by

$$\text{AVRB}(\widehat{F}) = \frac{B^{-1}}{11}\sum_{q=1}^{11}\left|\sum_{b=1}^{B}\frac{\widehat{F}(t_q)_b - F_y(t_q)}{F_y(t_q)}\right|, \quad \text{AVRE}(\widehat{F}) = \frac{1}{11}\sum_{q=1}^{11}\frac{MSE[\widehat{F}(t_q)]}{MSE[\widehat{F}_{HT}(t_q)]}$$

where b indexes the bth simulation run, $\widehat{F}$ is an estimator for the distribution function, $MSE[\widehat{F}(t)] = B^{-1}\sum_{b=1}^{B}[\widehat{F}(t)_b - F_y(t)]^2$ is the empirical mean square error for $\widehat{F}(t)$.

Table 1 provides the values AVRB and AVRE for the population with two non-response mechanism considered. From results, it is observed that the usual estimators $\widehat{F}_{HT}(t)$, $\widehat{F}_{CD}(t)$, $\widehat{F}_r(t)$ and $\widehat{F}_{RKM}(t)$ have a considerable bias for all sample sizes.The proposed calibration estimators significantly reduce the bias, especially the estimators $\hat{F}_{calTS}$, and the different versions of $\hat{F}_{calIKL1}$ and $\hat{F}_{calIKL2}$. In a similar way estimators $\widehat{F}_{CD}(t)$, $\widehat{F}_r(t)$ and $\widehat{F}_{RKM}(t)$ suffer an important loss in efficiency compared to the $\widehat{F}_{HT}(t)$ estimator for the raking mechanism. All the proposed calibration estimators exhibit greater efficiency than these estimators. $\hat{F}_{calTS}$, and the different versions of $\hat{F}_{calIKL1}$ and $\hat{F}_{calIKL2}$ show the best performance. There is no significant difference between the two versions of the estimators (raking and logit methods) in terms of efficiency although the raking method produces the estimators with fewer errors in most cases. There is no estimator that is uniformly better than the rest in terms of bias and error.

Tabla A1.1: Average relative bias (AVRB) and the average relative efficiency (AVRE) of compared estimators. The lowest value is denoted in bold

	Raking non-response mechanism									
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 125$		$n = 135$		$n = 145$		$n = 155$		$n = 165$	
$\widehat{F}_{HT}$	0.315	1.000	0.313	1.000	0.316	1.000	0.314	1.000	0.319	1.000
$\widehat{F}_{CD}$	0.367	2.444	0.372	2.695	0.370	2.807	0.372	3.013	0.367	2.796
$\widehat{F}_{r}$	0.359	2.318	0.363	2.502	0.360	2.623	0.360	2.725	0.358	2.525
$\widehat{F}_{RKM}$	0.339	2.242	0.344	2.431	0.340	2.535	0.344	2.717	0.340	2.515
$\widehat{F}_{calTS}$	**0.002**	0.270	**0.002**	0.256	0.002	0.247	0.007	0.243	**0.001**	0.211
$\widehat{F}_{calIDra}$	0.016	0.369	0.012	0.354	0.008	0.323	0.005	0.328	0.008	0.278
$\widehat{F}_{calIDlo}$	0.016	0.369	0.012	0.354	0.008	0.323	0.005	0.328	0.008	0.278
$\widehat{F}_{calIKL1ra}$	**0.002**	**0.264**	0.004	**0.254**	0.003	**0.241**	0.008	**0.238**	0.003	**0.210**
$\widehat{F}_{calIKL1lo}$	0.003	0.277	**0.002**	0.266	**0.001**	0.252	0.006	0.250	**0.001**	0.221
$\widehat{F}_{calIKL2ra}$	0.005	0.287	0.003	0.275	**0.001**	0.262	**0.004**	0.259	0.002	0.229
$\widehat{F}_{calIKL2lo}$	0.005	0.287	0.003	0.275	**0.001**	0.262	**0.004**	0.259	0.002	0.229
	Logistic non-response mechanism									
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 125$		$n = 135$		$n = 145$		$n = 155$		$n = 165$	
$\widehat{F}_{HT}$	0.340	1.000	0.341	1.000	0.337	1.000	0.342	1.000	0.340	1.000
$\widehat{F}_{CD}$	0.181	0.278	0.185	0.279	0.181	0.276	0.186	0.281	0.184	0.274
$\widehat{F}_{r}$	0.191	0.401	0.197	0.404	0.195	0.399	0.199	0.396	0.199	0.395
$\widehat{F}_{RKM}$	0.181	0.318	0.184	0.317	0.180	0.314	0.185	0.316	0.184	0.310
$\widehat{F}_{calTS}$	0.040	0.101	0.035	0.091	0.032	0.088	0.036	0.083	0.036	0.079
$\widehat{F}_{calIDra}$	**0.034**	0.102	**0.029**	0.093	**0.026**	0.089	**0.030**	0.084	**0.031**	0.079
$\widehat{F}_{calIDlo}$	0.035	0.103	0.030	0.093	0.027	0.089	0.031	0.084	0.032	0.080
$\widehat{F}_{calIKL1ra}$	0.036	**0.098**	0.030	**0.089**	0.029	**0.085**	0.033	**0.080**	0.032	**0.078**
$\widehat{F}_{calIKL1lo}$	0.037	**0.098**	0.031	**0.089**	0.030	0.086	0.034	**0.080**	0.033	0.079
$\widehat{F}_{calIKL2ra}$	0.036	**0.098**	0.031	**0.089**	0.030	**0.085**	0.033	**0.080**	0.032	**0.078**
$\widehat{F}_{calIKL2lo}$	0.038	0.099	0.032	**0.089**	0.031	0.086	0.034	0.081	0.033	0.079

7. Conclusion

This paper describes how calibration weighting can be used to adjust the design weights to increase the efficiency of finite distribution function for a sample survey when there is unit nonresponse. We propose two calibration methods to reduce the non-response bias. The first method is based on two-step calibration weighting. The first calibration is designed to remove the non-response bias. The second one to decrease the sampling error in the estimation of the distribution function. This method allows different variables to be used in each phase, since the model for non-response and the predictive model can be very different. This estimator given by (1.4) is computationally simple. The last method is based on model and calibration variables. The calibration is done in a single stage, but different variables are also used to model the lack of response and for the calibration equation. Different model are also proposed to model the nonresponse. The problem with this methodology is the difficulty in solving the calibration equation. Various iterative methods are proposed to obtain the weights.

Our limited simulation study clearly shows the gain in reduction of bias and precision achieved when calibration is used for nonresponse that is not missing completely at random. There is no estimator that is uniformly better than the rest in terms of bias and error. The $\hat{F}_{calIKL1}$ and $\hat{F}_{calIKL2}$ estimators produce the best estimates in terms of the least error in most cases. The computational simplicity of the estimator in two stages $\hat{F}_{calTS}$ is noteworthy.

Acknowledgements

This work is partially supported by Ministerio de Economía y Competitividad of Spain (grant MTM2015-63609-R).

References

[1] Beaumont, J. F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(3), 445-458.

[2] Chambers, R.L., & Dunstan, A. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

[3] Chang, T., & Kott, P. S. (2008). Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. *Biometrika*, 95, 557-571.

[4] Deville,J.C., & Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

[5] Deville, J. C. (2000). Generalized Calibration and Application to Weighting for Non-response. *COMPSTAT: Proceedings in Computational Statistics, 14th Symposium, Utrecht, The Netherlands*, eds. J. G. Bethlehem and P. G. M. van der Heijden, New York: Springer-Verlag, 65-76.

[6] Harms, T., & Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32, 37–52.

[7] Kott, P. S., & Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. *Survey Research Methods*, 6(2), 105-111.

[8] Kott, P.S., & Liao, D. (2015). One step or two? Calibration weihting form a complete list frame with nonresponse. *Survey Methodology*, 41(1), 165-181.

[9] Kott, P.S., & Liao, D. (2017). Calibration weighting for nonresponse that is not missing at random: allowing more calibration than response-model variables. *Journal of Survey Statistics and Methodology*, 5, 159–174.

[10] Lesage,E., Haziza, D., & D'Haultfoeuille. (2019). A Cautionary Tale on Instrumental Calibration for the Treatment of Nonignorable Unit Nonresponse in Surveys. Journal of the American Statistical Association, 114(526), 906-915.

[11] Lundström, S., & Särndal, C. E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

[12] Martínez, S., Rueda, M., Arcos, A., & Martínez, H. (2010). Optimum calibration points estimating distribution functions. *Journal of Computational and Applied Mathematics*, 233, 2265-2277.

[13] Rao, J.N.K., Kovar, J.G., & Mantel, H.J. (1990). On estimating distribution function and quantiles from survey data using auxiliary information. *Biometrika*, 77(2), 365-375.

[14] Rota, B., & Laitila,T. (2015). Comparisons of some weigting methods for nonresponse adjustment. *Lithuanian Journal of Statistics*, 54(1), 69-83.

[15] Rueda, M., Martínez, S., Martínez, H., & Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137(2), 435-448.

[16] Särndal, C.E., & Lundström, S. *Estimation in Surveys with Nonresponse*. John Wiley & Sons.

[17] Tille, Y., & Matei, A. (2021) sampling: Survey Sampling. A software routine available online at https://CRAN.R-project.org/package=sampling

Apéndice A2

Calibration adjustment for dealing with non-response in the estimation of poverty measures

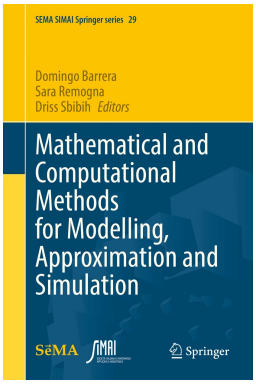
Illescas, María Dolores; Martínez, Sergio; Rueda, María del Mar; Arcos, Antonio (2022)

Calibration adjustment for dealing with nonresponse in the estimation of poverty measures.

En *Barrera, D., Remogna, S., Sbibih, D. (eds) Mathematical and Computational Methods for Modelling, Approximation and Simulation. SEMA SIMAI Springer Series*, Vol. 29, pp. 209–236.

Springer, Cham.

DOI: 10.1007/978-3-030-94339-4_11



Ranking General Editoriales			
SPI Year	ICEE	Rank	Quartile
2018	33.061	4/259	Q1

Abstract

The analysis of poverty measures has been receiving increased attention in recent years. This paper contributes to the literature by developing percentile ratio estimators when there are missing data. Calibration adjustment is used for treating the non-response bias. Variances of the proposed estimators could be not expressible by simple formulae and resampling techniques are investigated for obtaining variance estimators. A numerical example based on data from the Spanish Household Panel Survey is taken up to illustrate how suggested procedures can perform better than existing ones.

1. Introduction

The analysis of poverty measures is a topic of increased interest to society. The official poverty rate and the number of people in poverty are important measures of the country's economic wellbeing. The common characteristic of many poverty measures is their complexity. The literature on survey sampling is usually focused on the goal of estimating linear parameters. However when the variable of interest is a measure of wages or income, the distribution function is a relevant tool because is required to calculate the poverty line, the low income proportion, the poverty gap and other poverty measures.

The lack of response is a growing problem in economic surveys. Although there are many procedures for their treatment, few efficient techniques have been developed for their treatment in the estimation of non-linear parameters. Recently Rueda et al. (2021) have proposed various estimators for the distribution function in the presence of missing data. Using these estimators, we first propose new estimators for several poverty measures, which efficiently use auxiliary information at the estimation stage. Due to the complexity of the percentile ratios and the complex sampling designs used by the official sample surveys, variances of these complex statistics could be not expressible by simple formulae. Additional techniques for variance estimation are therefore required under this scenario.

This paper is organized as follows. Section 2 introduces the estimation of the distribution function when there are missing data. In Sect.3, the proposed percentile ratio estimators are described. In Sect.4 we derive resampling techniques for the problem of the variance estimation of percentile ratio estimators. A simulation study based on data derived from the Spanish Household Panel Survey is presented in Sect.5. This study shows how the proposed estimates of the poverty measures perform better than usual calibration estimators in reduction of bias and precision when calibration is used for nonresponse that is not missing at random.

2. Calibrating the distribution function for treating the non-response

Consider a finite population $U = \{1, \ldots, N\}$ consisting of N different and identifiable units. Let us assume a sampling design d defined in U with positive first-order inclusion probabilities $\pi_i \; i, \in U$. Let $d_i = \pi_i^{-1}$ denote the sampling design-basic weight for unit $i \in U$ which is known. We assume missing data on the sample s obtained by the sampling design d. Let us denote by s_r, the respondent sample of size r, and s_m the non-respondent sample of size $n - r$.

Let y_i be the value of the character under study. The distribution function $F_y(t)$ can be estimated by the Horvitz-Thompson estimator:

$$\widehat{F}_{HT}(t) = \frac{1}{N} \sum_{k \in s_r} d_k \Delta(t - y_k) \tag{2.1}$$

where

$$\Delta(t - y_k) = \begin{cases} 0 & \text{if } t < y_k \\ 1 & \text{if } t \geq y_k \end{cases}$$

and $d_k = 1/\pi_k$, the basic design weights.

This estimator is biased for the distribution function. There are several approach for dealing with nonresponse. The most important method is weighting. We assume the existence of auxiliary information relative to several variables related to the main variable y, $\mathbf{x} = (x_1, x_2, \ldots, x_J)'$. Based on this auxiliary information, calibration weighting is used in Rueda et al. (2021) to propose three methods to reduce the non-response bias in the estimation of the distribution function:

- The first method is based on the methodology proposed in Rueda et al. (2007a). We define a pseudo-variable $g_k = \widehat{\beta}' \mathbf{x}_k$ for $k = 1, 2, ...N$, where $\widehat{\beta} = \left(\sum_{j \in s_r} d_j \mathbf{x}_j \mathbf{x}_j^{'} \right)^{-1} \cdot \sum_{j \in s_r} d_j \mathbf{x}_j y_j$.

  Thus we define a calibrated estimator by imposing that the calibrated weights w_k evaluated in the observed sample give perfect estimates for the distribution function in a set of predetermined points t_j for $j = 1, 2, \ldots, P$ that we choose arbitrarily:

$$\frac{1}{N} \sum_{k \in s_r} w_k \Delta(\mathbf{t} - g_k) = F_g(\mathbf{t}) \tag{2.2}$$

  where $F_g(\mathbf{t})$ denotes the finite distribution function of the pseudo-variable g_k evaluated at the point $\mathbf{t} = (t_1, ..., t_p)'$ and $\Delta(\mathbf{t} - g_k) = (\Delta(t_1 - g_k), ..., \Delta(t_p - g_k))'$.

A common way to compute calibration weights is linearly (using the chi-square distance method) and we obtain an explicit expression of the estimator as:

$$\hat{F}_{cal}^{(1)}(t) = \frac{1}{N}\sum_{k\in s_r} w_k^{(1)}\Delta(t-y_k) =$$

$$\widehat{F}_{HT}(t) + \left(F_g(\mathbf{t}) - \frac{1}{N}\sum_{k\in s_r} d_k\Delta(\mathbf{t}-g_k)\right)' \cdot T^{-1} \cdot \sum_{k\in s_r} d_k\Delta(\mathbf{t}-g_k)\Delta(t-y_k) \tag{2.3}$$

where $T = \sum_{k\in s_r} d_k\Delta(\mathbf{t}-g_k)\Delta(\mathbf{t}-g_k)'$.

Following Rueda et al. (2007a), if we denote by $k_i = \sum_{k\in s_r} d_k\Delta(t_i-g_k)$ for $i = 1,\ldots,P$ the condition $k_i > k_{i-1}$ for $i = 2,\ldots,P$ guarantees the existence of T^{-1}.

- The second method is based on two-step calibration weighting as in the work of (Kott & Liao (2015)):

  1.- The first calibration is designed to remove the non-response bias.

  Consider the M vector of explanatory model variables, $\mathbf{x}_k^*$ which population totals $\sum_U \mathbf{x}_k^*$ are know. The calibration under the restrictions $\sum_{s_r} v_k^{(1)}\mathbf{x}_k^* = \sum_U \mathbf{x}_k^*$ yields the calibrations weights $v_k^{(1)}, k = 1,\ldots,s_r$.

  2.- The second one to decrease the sampling error in the estimation of the distribution function.

  The auxiliary information of the calibration variables $\mathbf{x}$ is incorporated through the calibrated weights $v_k^{(2)}$ obtained with the restrictions $\sum_{s_r} v_k^{(2)}\Delta(\mathbf{t}-g_k) = F_g(\mathbf{t})$. The final estimator is given by

$$\hat{F}_{cal}^{(2)}(t) = \frac{1}{N}\sum_{k\in s_r} w_k^{(2)}\Delta(t-y_k) = \frac{1}{N}\sum_{k\in s_r} v_k^{(2)}v_k^{(1)}\Delta(t-y_k) \tag{2.4}$$

  This method allows different variables to be used in each phase (model variables $\mathbf{x}_k^*$ and calibration variables $\mathbf{x}$), since the model for non-response and the predictive model can be very different.

- The last method is based on instrumental variables (Deville (2000);Kott & Liao (2017)). The calibration is done in a single stage, but different variables are also used to model the lack of response and for the calibration equation. By assuming that the probability of response can be modeled by: $\theta_k = f(\gamma'\mathbf{x}_k^*)$ for some vector parameter γ, where $h(\cdot) = 1/f(\cdot)$ is a known and everywhere monotonic and twice differentiable function. We denote as $\mathbf{z}_k = \Delta(\mathbf{t}-g_k)$

  The calibration equation is given by:

$$\frac{1}{N}\sum_{k\in s_r}\frac{d_k}{f(\hat{\gamma}'\mathbf{x}_k^*)}\mathbf{z}_k = \frac{1}{N}\sum_{k\in s_r} d_k h(\hat{\gamma}'\mathbf{x}_k^*)\mathbf{z}_k = F_g(\mathbf{t}) \tag{2.5}$$

and the resulting calibrated estimator is:

$$\hat{F}_{cal}^{(3)}(t) = \frac{1}{N}\sum_{k\in s_r} w_k^{(3)}\Delta(t-y_k) = \frac{1}{N}\sum_{k\in s_r} d_k h(\hat{\gamma}'\mathbf{x}_k^*)\Delta(t-y_k) \tag{2.6}$$

where $\hat{\gamma}$ is a consistent estimator of vector γ. Authors use several approximation methods for deriving the solution of the minimization problem. We denote by $\widehat{F}_{Dcal}^{(3)}(t)$ the estimator based in the Deville's approach Deville (2000) which needs to meet the condition $M = P$. To consider more calibration restrinctions in equation (2) than M, we consider the estimator $\widehat{F}_{KL1cal}^{(3)}(t)$ and $\widehat{F}_{KL2cal}^{(3)}(t)$ based on Kott & Liao (2017) where $P > M$.

3. Poverty measures estimation with missing values

Currently, poverty measurement, wage inequality, inequality and life condition are overriding issues for governments and society. Some indices and poverty measures used in the poverty evaluation and income inequality measurement are based on quantile and quantiles ratios. Thereby, Eurostat currently set the poverty line (the population threshold for classification into poor and nonpoor) equal to sixty percent of the equivalized net income median Q_{50}. On the other hand, the percentile ratios Q_{95}/Q_{20}; Q_{90}/Q_{10} and Q_{80}/Q_{20} (Jones & Weinberg (2000)); Q_{95}/Q_{50} and Q_{50}/Q_{10} (Machin et al. (2003), Burtless (1999)); Q_{50}/Q_5 and Q_{50}/Q_{25} (Dickens & Manning (2004)) have been considered as measures for wage inequality. We focus on estimating the poverty measures based on percentile ratios.

The population α-quantile of y is defined as follows

$$Q_y(\alpha) = \text{ínf}\{t : F_y(t) \geq \alpha\} = F_y^{-1}(\alpha) \tag{2.7}$$

A general procedure to incorporate the auxiliary information in the estimation of $Q_y(\alpha)$ is based on the obtainment of an indirect estimator $\widehat{F}_y(t)$ of $F_y(t)$ that fulfills the distribution function's properties. Under this assumption, the quantile $Q_y(\alpha)$ can be estimated in a following way:

$$\widehat{Q}_y(\alpha) = \text{ínf}\{t : \widehat{F}_y(t) \geq \alpha\} = \widehat{F}_y^{-1}(\alpha) \tag{2.8}$$

The distribution function estimators described in the previous section allow us to incorporate the auxiliary information in the estimation of quantiles in the presence of non-response and obtain estimators for percentile ratios. Perhaps, some of these calibrated estimators do not satisfy all the properties of the distribution function and consequently for its application in the estimation of quantiles some modifications are necessary. Specifically, the properties that an estimator $\widehat{F}_y(t)$ of the distribution function $F_y(t)$ must meet are the following:

i. $\widehat{F}_y(t)$ is continuous on the right.

ii. $\widehat{F}_y(t)$ is monotone nondecreasing,

iii. a) $\lim_{t\to-\infty} \widehat{F}_y(t) = 0$ and b) $\lim_{t\to+\infty} \widehat{F}_y(t) = 1$.

Firstly, it's easy to see that all estimators satisfy the conditions (i) and iii.(a). Secondly, following Rueda et al. (2007a), the estimator $\widehat{F}^{(1)}_{cal}(t)$ meet the rest of conditions if t_P is sufficiently large (i.e $F_g(t_P) = 1$). On the other hand, it's easy to see that $\widehat{F}^{(2)}_{cal}(t)$ satisfy the condition iii.(b) if t_P is sufficiently large but it is not monotone nondecreasing in general. Thus, we can apply the procedure described in Rao et al. (1990). This procedure, for a general estimator $\widehat{F}_y$, is defined in the following way:

$$\tilde{F}_y(y_{[1]}) = \widehat{F}_y(y_{[1]}), \quad \tilde{F}_y(y_{[i]}) = \text{máx}\{\widehat{F}_y(y_{[i]}),\, \tilde{F}_y(y_{[i-1]})\} \quad i = 2, \ldots, r \tag{2.9}$$

Finally, all estimators based on $\widehat{F}^{(3)}_{cal}(t)$ are nondecreasing if $\theta_k = f(\gamma' \mathbf{x}^*_k) \geq 0$ for all $k \in U$ (response model based on logit, raking and logistic methods) because the calibration weights $\omega^{(3)}_k \geq 0$. Moreover, $\widehat{F}^{(3)}_{Dcali}(t)$ fulfills condition iii.(b) with t_P sufficiently large whereas following Rueda et al. (2021), $\widehat{F}^{(3)}_{KL1cali}(t)$ and $\widehat{F}^{(3)}_{KL2cali}(t)$ meet condition iii.(b) if in addition to considering t_P sufficiently large, a component of the vector $\mathbf{x}^*_k$ contains all 1's.

Based on the population distribution function $F_y(t)$, given two values $1 > \alpha_1 > \alpha_2 > 0$, the percentile ratio $R(\alpha_1, \alpha_2)$ is define as follow:

$$R(\alpha_1, \alpha_2) = \frac{Q_y(\alpha_1)}{Q_y(\alpha_2)} \tag{2.10}$$

and it can be estimated with a generic quantile estimator $\widehat{Q}_y(\alpha)$ as follows:

$$\widehat{R}(\alpha_1, \alpha_2) = \frac{\widehat{Q}_y(\alpha_1)}{\widehat{Q}_y(\alpha_2)} \tag{2.11}$$

Thus, the quantile estimator derived from $\widehat{F}^{(1)}_{cal}$; $\widehat{F}^{(2)}_{cal}$ and $\widehat{F}^{(3)}_{cal}$ can be employed in the estimation of $R(\alpha_1, \alpha_2)$.

4. Variance estimation for percentile ratio estimators with resampling method

Given the complexity of the proposed percentile ratio estimators, we have considered the use of bootstrap techniques for estimating variance and developing confidence intervals associated with the proposed calibration estimators. In this study, we consider the frameworks proposed in Booth et al. (1994), Antal & Tillé (2011) and Antal & Tillé (2014).

First, the bootstrap procedure described in Booth et al. (1994) consider the repetition of sample units for creating artificial bootstrap populations. The bootstrap samples are drawing with the original sampling design from artificial populations. Specifically, if the population size $N = n \cdot q + m$ with $0 < m < n$, the artificial population U_B is obtained with q repetitions of s and an additional sample of size m selected by simple random sampling without replacement from s. Given a generic percentile ratio estimator $\widehat{R}(\alpha_1, \alpha_2)$, if we consider M independent artificial populations U_B^j with $j = 1, \ldots, M$ and for each pseudo population U_B^j we select K bootstrap samples $s_1^j, \ldots, s_K^j$ with sample size n, we can compute the bootstrap estimates $\widehat{R}^*(\alpha_1, \alpha_2)_h^j$ with the sample s_h^j for the population U_B^j and following (Chauvet et al. (2007)), we can compute:

$$\widehat{V}_j = \frac{1}{K-1} \sum_{h=1}^{K} (\widehat{R}^*(\alpha_1, \alpha_2)_h^j - \widehat{R}_j^*(\alpha_1, \alpha_2))^2 \tag{2.12}$$

where

$$\widehat{R}_j^*(\alpha_1, \alpha_2) = \frac{1}{K} \sum_{h=1}^{K} \widehat{R}^*(\alpha_1, \alpha_2)_h^j \tag{2.13}$$

Finally, the variance estimation for the estimator $\widehat{R}(\alpha_1, \alpha_2)$ is given by:

$$\widehat{V}(\widehat{R}(\alpha_1, \alpha_2)) = \frac{1}{M} \sum_{j=1}^{M} \widehat{V}_j \tag{2.14}$$

On the other hand, Antal & Tillé (2011) and Antal & Tillé (2014) have proposed a direct bootstrap methods where it is not necessary to obtain an artificial population, since the bootstrap samples are drawn from s under a sampling scheme different from the original sampling design. Both frameworks (Antal & Tillé (2011) and Antal & Tillé (2014)) can be applied under several sample designs, but particularly, if the sample s is drawing with simple random sampling without replacement, the sampling design proposed by Antal & Tillé (2011) select two samples from s, the first one is drawing by simple random sampling without replacement

and the second one is drawing with one-one sampling design (a sampling design for resampling). Similarly, under simple random sampling without replacement, the sampling design proposed by Antal & Tillé (2014) draw a first sample with Bernoulli design and a second sample with double half sampling design (another sampling design for resampling). For more details see Antal & Tillé (2011) and Antal & Tillé (2014).
For two frameworks, given a percentile ratio estimator $\widehat{R}(\alpha_1, \alpha_2)$, we draw M bootstrap samples $s_1^*, \ldots, s_M^*$ from s, according to the sampling schemes of Antal & Tillé (2011) and Antal & Tillé (2014) respectively. The bootstrap estimation for variance of the estimator $\widehat{R}(\alpha_1, \alpha_2)$ is given by:

$$\widehat{V}(\widehat{R}(\alpha_1, \alpha_2)) = \frac{1}{M} \sum_{j=1}^{M} (\widehat{R}(\alpha_1, \alpha_2)_j^* - \bar{R}(\alpha_1, \alpha_2)^*)^2 \tag{2.15}$$

where $\widehat{R}(\alpha_1, \alpha_2)_j^*$ is the bootstrap estimator computed with the bootstrap sample s_j^* and

$$\bar{R}(\alpha_1, \alpha_2)^* = \frac{1}{M} \sum_{j=1}^{M} \widehat{R}(\alpha_1, \alpha_2)_j^* \tag{2.16}$$

Finally, based on the variance estimation $\widehat{V}(\widehat{R}(\alpha_1, \alpha_2))$ obtained with a bootstrap method, the $1 - \alpha$ level confidence interval based on the approximation by a standard normal distribution is defined as follows:

$$\left[\widehat{R}(\alpha_1, \alpha_2) - z_{1-\alpha/2} \cdot \sqrt{\widehat{V}(\widehat{R}(\alpha_1, \alpha_2))}, \widehat{R}(\alpha_1, \alpha_2) + z_{1-\alpha/2} \cdot \sqrt{\widehat{V}(\widehat{R}(\alpha_1, \alpha_2))}\right] \tag{2.17}$$

where z_α is the α quantile of the standard normal distribution. For all bootstrap methods included in this study, we can compute with this procedure the respective confident interval.

5. Simulation study

To determine the behaviour of the estimators when they are applied to real data we consider data from the region of Andalusia of 2016 Spanish living conditions survey carried out by the Instituto Nacional de Estadística (INE) of Spain. The survey data collected are considered as a population with size $N = 1442$ and samples are selected from it. The study variable y is the equivalised net income and the auxiliary variables included are the following dummy variables b_1 ="Home without mortgage", b_2 ="Four-bedroom home" and b_3 ="Can the home afford to go on vacation away from home, at least one week a year?". We considered the vector of model variables $(x_k^*)' = (1, b_{1k})$ and the vector of calibration variables $(x_k)' = (1, b_{1k}, b_{2k}, b_{3k})$.

We consider four response mechanism where the probability of the k-th individual of responde is given

by

$$\theta_k = \frac{1}{exp(A + b_{1k}/B)} \tag{2.18}$$

with different values for A and B.

The ratio estimators considered in this simulation study, based on the respondent sample s_r, are obtained from the Horvitz-Thompson estimator $\widehat{F}_{HT}(t)$. We denoted by $\widehat{R}_D^{(3)}$ the calibration estimator based on (Deville (2000)) and we denoted by $\widehat{R}_{KL1}^{(3)}$ and $\widehat{R}_{KL2}^{(3)}$ the calibration estimators based on (Kott & Liao (2017)). The estimator $\widehat{R}_{cal}^{(1)}$ has been included only with comparative purposes with respect to the rest of proposed estimators because it only considers the respondent sample and it does not deal with nonresponse whereas the rest of the estimators proposed try to deal with the bias produced by nonresponse. Although the real response mechanism considered is based on raking method, for $\widehat{R}_D^{(3)}$, $\widehat{R}_{KL1}^{(3)}$ and $\widehat{R}_{KL2}^{(3)}$ three versions of them are computed based on linear, raking and logit (l; u) response models.

We selected $W = 1000$ samples with several sample sizes, $n = 100$, $n = 125$, $n = 150$ and $n = 200$, under simple random sampling without replacement (SRSWOR) and for each estimator included in the simulation study, we computed estimates of $R(\alpha_1, \alpha_2)$ for 50th/25th, 80th/20th, 90th/10th, 90th/20th, 95th/20th and 95th/50th. The performance of each estimator is measured by the relative bias, (RB), and the relative efficiency (RE), given by

$$\mathrm{RB}(\widehat{R}(\alpha_1, \alpha_2)) = \sum_{w=1}^{W} \frac{\left(\widehat{R}(\alpha_1, \alpha_2)\right)_w - R(\alpha_1, \alpha_2)}{R(\alpha_1, \alpha_2)} \tag{2.19}$$

$$\mathrm{RE}(\widehat{R}(\alpha_1, \alpha_2)) = \frac{\sum_{w=1}^{W} \left[\left(\widehat{R}(\alpha_1, \alpha_2)\right)_w - R(\alpha_1, \alpha_2)\right]^2}{\sum_{w=1}^{W} \left[\left(\widehat{R}_{HT}(\alpha_1, \alpha_2)\right)_w - R(\alpha_1, \alpha_2)\right]^2}, \tag{2.20}$$

where $\widehat{R}(\alpha_1, \alpha_2)$ is a percentile ratio estimator and $\widehat{R}_{HT}(\alpha_1, \alpha_2)$ is the percentile ratio estimator based in the Horvitz-Thompson $\widehat{F}_{HT}(t)$ estimator .

Tables A2.1-A2.6 provide the values of the relative bias and the relative efficiency for this population for several sample sizes and response mechanism of the estimators compared.

Tabla A2.1: RB and RE for several sample sizes of the estimators of $R(0{,}5, 0{,}25)$. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY.

$1/\exp(0{,}5 + b_1/3)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	1.4641	1	1.6163	1	1.7126	1	1.932	1
$\widehat{R}^{(1)}_{cal}$	0.0543	0.0049	0.0509	0.0032	0.0481	0.0024	0.0461	0.0014
$\widehat{R}^{(2)}_{cal}$	0.0107	0.0038	0.0056	0.0022	0.0023	0.0016	-0.001	9e-04
$\widehat{R}_{KL1cali}$	0.0116	0.0035	0.0051	0.0021	0.0019	0.0015	-8e-04	8e-04
$\widehat{R}_{KL1calr}$	0.0109	0.0035	0.0048	0.0021	0.0021	0.0015	-0.0011	9e-04
$\widehat{R}_{KL1calo}$	0.011	0.0036	0.0048	0.0021	0.002	0.0015	-0.0012	8e-04
$\widehat{R}_{KL2cali}$	0.0117	0.0035	0.005	0.0021	0.002	0.0015	-8e-04	8e-04
$\widehat{R}_{KL2calr}$	0.0108	0.0035	0.0046	0.0021	0.0019	0.0015	-0.0012	8e-04
$\widehat{R}_{KL2calo}$	0.0108	0.0035	0.0046	0.0021	0.0019	0.0015	-0.0012	8e-04
$\widehat{R}_{Dcali}$	0.0101	0.0036	0.0043	0.0021	0.0011	0.0016	-0.001	9e-04
$\widehat{R}_{Dcalr}$	0.0092	0.0036	0.004	0.0021	0.0011	0.0016	-0.0013	9e-04
$\widehat{R}_{Dcalo}$	0.0092	0.0036	0.004	0.0021	0.0011	0.0016	-0.0013	9e-04
$1/\exp(0{,}25 + b_1/2)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.5152	1	0.502	1	0.4624	1	0.4441	1
$\widehat{R}^{(1)}_{cal}$	0.046	0.0264	0.041	0.0236	0.0391	0.0249	0.0385	0.0266
$\widehat{R}^{(1)}_{cal}$	0.0074	0.0204	0.0024	0.0176	-0.0018	0.0175	-0.0019	0.0172
$\widehat{R}_{KL1cali}$	0.0072	0.0188	0.003	0.0168	-0.0014	0.0173	-0.0014	0.0169
$\widehat{R}_{KL1calr}$	0.0069	0.0187	0.0026	0.0168	-0.0015	0.0173	-0.0017	0.0169
$\widehat{R}_{KL1calo}$	0.0068	0.0188	0.0026	0.0168	-0.0015	0.0172	-0.0016	0.0169
$\widehat{R}_{KL2cali}$	0.0072	0.0188	0.0029	0.0168	-0.0014	0.0173	-0.0014	0.0169
$\widehat{R}_{KL2calr}$	0.0067	0.0188	0.0026	0.0168	-0.0014	0.0173	-0.0016	0.0169
$\widehat{R}_{KL2calo}$	0.0067	0.0188	0.0026	0.0168	-0.0014	0.0173	-0.0016	0.0169
$\widehat{R}_{Dcali}$	0.0063	0.0196	0.0025	0.0174	-0.0018	0.0175	-0.0016	0.0171
$\widehat{R}_{Dcalr}$	0.0056	0.0196	0.0023	0.0174	-0.0018	0.0175	-0.0019	0.0171
$\widehat{R}_{Dcalo}$	0.0056	0.0196	0.0023	0.0174	-0.0018	0.0175	-0.0019	0.0171
$1/\exp(0{,}75 + b_1/8)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	1.7102	1	1.9192	1	2.1328	1	2.4505	1
$\widehat{R}^{(1)}_{cal}$	0.0601	0.0046	0.0539	0.0028	0.0491	0.002	0.0467	0.0012
$\widehat{R}^{(2)}_{cal}$	0.0111	0.0033	0.0066	0.0021	0.0039	0.0014	0.0015	7e-04
$\widehat{R}_{KL1cali}$	0.0132	0.0031	0.0073	0.002	0.003	0.0013	6e-04	7e-04
$\widehat{R}_{KL1calr}$	0.0121	0.0031	0.0067	0.002	0.0027	0.0013	2e-04	7e-04
$\widehat{R}_{KL1calo}$	0.0121	0.0031	0.0069	0.002	0.0028	0.0013	3e-04	7e-04
$\widehat{R}_{KL2cali}$	0.013	0.0031	0.0072	0.002	0.0028	0.0013	6e-04	7e-04
$\widehat{R}_{KL2calr}$	0.0121	0.0031	0.0069	0.002	0.0028	0.0013	3e-04	7e-04
$\widehat{R}_{KL2calo}$	0.0121	0.0031	0.0069	0.002	0.0028	0.0013	3e-04	7e-04
$\widehat{R}_{Dcali}$	0.0105	0.0031	0.0063	0.002	0.0019	0.0013	1e-04	7e-04
$\widehat{R}_{Dcalr}$	0.0095	0.0031	0.0061	0.002	0.0019	0.0013	-3e-04	7e-04
$\widehat{R}_{Dcalo}$	0.0095	0.0031	0.0061	0.002	0.0019	0.0013	-3e-04	7e-04
$1/\exp(0{,}125 + b_1/2)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.2424	1	0.2331	1	0.2224	1	0.2312	1
$\widehat{R}^{(1)}_{cal}$	0.039	0.1481	0.0381	0.1505	0.0339	0.1305	0.0341	0.1001
$\widehat{R}^{(1)}_{cal}$	0.0042	0.1224	0.0021	0.1145	-0.0014	0.0951	-0.0013	0.0679
$\widehat{R}_{KL1cali}$	0.0058	0.1182	0.0019	0.1096	-0.0017	0.0917	-0.0011	0.0668
$\widehat{R}_{KL1calr}$	0.0052	0.1159	0.0022	0.1094	-0.0018	0.0913	-0.0013	0.0667
$\widehat{R}_{KL1calo}$	0.0052	0.1158	0.0021	0.1094	-0.0019	0.0912	-0.0013	0.0667
$\widehat{R}_{KL2cali}$	0.0057	0.1178	0.0019	0.1096	-0.0017	0.0917	-0.0011	0.0667
$\widehat{R}_{KL2calr}$	0.0053	0.1177	0.0018	0.1097	-0.0017	0.0917	-0.0013	0.0668
$\widehat{R}_{KL2calo}$	0.0053	0.1177	0.0018	0.1097	-0.0017	0.0917	-0.0013	0.0668
$\widehat{R}_{Dcali}$	0.0056	0.1196	0.0014	0.1132	-0.0022	0.0945	-0.0014	0.0676
$\widehat{R}_{Dcalr}$	0.0051	0.1195	0.0013	0.1133	-0.0022	0.0945	-0.0016	0.0676
$\widehat{R}_{Dcalo}$	0.0051	0.1195	0.0013	0.1133	-0.0022	0.0945	-0.0016	0.0676

Results from Tables A2.1-A2.6 show an important bias for the estimator $\widehat{R}_{HT}$ in almost percentile ratios. The estimator $\widehat{R}^{(1)}_{cal}$ is not capable of correcting the bias in several situations, giving worse estimates than the HT estimator for some ratios and some response mechanisms. The proposed estimators have better values of RB with slight differences between them, although there is no uniformly better estimator than the rest. Regarding efficiency, in general, the proposed estimators show the best performance for all sample sizes. Finally, in terms of bias and efficiency, there are no differences between the three versions of the estimators (linear method, raking and logit (l; u)) for the estimators $\widehat{R}^{(3)}_{cal}$.

For the variance estimation and confidence intervals, we computed the coverage probability (CP) , the lower (L) and the upper (U) tail error rates of the 95 % confidence intervals, in percentage and the average length (AL) of the confidence intervals for each estimator and each bootstrap method.
Concerning the variance estimation and confidence intervals, we used 1,000 bootstrap replications from each initial sample with all bootstrap methods included in the study to compute CP, L, U and AL of the 95 % confidence intervals for each percentile ratio estimator considered. Result from this simulation study for some percentile ratios are presented in Tables A2.7-A2.9.

From bootstrap estimates, it is observed that:

- All three bootstrap methods produce intervals with high true coverage.

- None of the intervals constructed with each estimator have problems of lack of coverage.

- The intervals obtained from the proposed calibration estimators always provide intervals with less amplitude than the intervals obtained from $\widehat{R}_{HT}$ and $\widehat{R}^{(1)}_{cal}$.

- The last method (Antal & Tillé (2011)) provides results very similar to the (Antal & Tillé (2014)) method.

6. Conclusion

In this study we use calibration techniques to estimate poverty measures based on percentiles ratios in presence of missing data through a more efficient estimation of the distribution function. The simulation study included shows the improvement in bias and efficiency with the two proposed calibration techniques, $\widehat{R}^{(2)}_{cal}$ and $\widehat{R}^{(3)}_{cali}$. The first one is based in two-step calibration method (Kott & Liao (2015)). In the first step, the weighting is designed to remove the non-response bias while in the second step the weighting is designed to decrease the sampling error in the estimation of the distribution function. The second method is based on calibration weighting with instrumental variables (Kott & Liao (2017)).

Tabla A2.2: RB and RE for several sample sizes of the estimators of $R(0{,}8, 0{,}2)$. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY.

$1/\exp(0{,}5 + b_1/3)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	1.0708	1	1.2001	1	1.3566	1	1.5867	1
$\widehat{R}_{cal}^{(1)}$	0.3746	0.1456	0.3339	0.0729	0.3264	0.055	0.328	0.0384
$\widehat{R}_{cal}^{(2)}$	0.036	0.0174	0.0212	0.0087	0.0183	0.0063	0.0166	0.0035
$\widehat{R}_{KL1cali}$	0.0375	0.0169	0.0225	0.0084	0.0189	0.0061	0.0176	0.0035
$\widehat{R}_{KL1calr}$	0.0372	0.0167	0.0229	0.0085	0.019	0.0061	0.0172	0.0034
$\widehat{R}_{KL1calo}$	0.0372	0.0167	0.0224	0.0085	0.019	0.0061	0.0174	0.0034
$\widehat{R}_{KL2cali}$	0.0374	0.0168	0.0224	0.0084	0.0189	0.0061	0.0177	0.0035
$\widehat{R}_{KL2calr}$	0.0374	0.0169	0.0224	0.0085	0.0189	0.0061	0.0176	0.0034
$\widehat{R}_{KL2calo}$	0.0374	0.0169	0.0224	0.0085	0.0189	0.0061	0.0176	0.0034
$\widehat{R}_{Dcali}$	0.0371	0.0172	0.0232	0.0088	0.0182	0.0063	0.0179	0.0035
$\widehat{R}_{Dcalr}$	0.0372	0.0173	0.0233	0.0088	0.0182	0.0063	0.0178	0.0035
$\widehat{R}_{Dcalo}$	0.0372	0.0173	0.0233	0.0088	0.0182	0.0063	0.0178	0.0035
$1/\exp(0{,}25 + b_1/2)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	1.3582	1	1.5555	1	1.7488	1	1.935	1
$\widehat{R}_{cal}^{(1)}$	0.3003	0.0607	0.2905	0.036	0.2966	0.0289	0.2842	0.0209
$\widehat{R}_{cal}^{(2)}$	0.0192	0.0083	0.0252	0.0055	0.021	0.0034	0.0155	0.0021
$\widehat{R}_{KL1cali}$	0.0204	0.0084	0.0243	0.0052	0.0217	0.0032	0.0155	0.0021
$\widehat{R}_{KL1calr}$	0.0201	0.0082	0.0243	0.0052	0.0213	0.0032	0.0156	0.0021
$\widehat{R}_{KL1calo}$	0.0206	0.0083	0.0239	0.0052	0.0215	0.0032	0.0156	0.0021
$\widehat{R}_{KL2cali}$	0.0205	0.0084	0.0241	0.0052	0.0216	0.0032	0.0155	0.0021
$\widehat{R}_{KL2calr}$	0.0201	0.0083	0.0243	0.0052	0.0216	0.0032	0.0156	0.0021
$\widehat{R}_{KL2calo}$	0.0201	0.0083	0.0243	0.0052	0.0216	0.0032	0.0156	0.0021
$\widehat{R}_{Dcali}$	0.0197	0.0088	0.0232	0.0053	0.0222	0.0033	0.0155	0.0021
$\widehat{R}_{Dcalr}$	0.0195	0.0088	0.0231	0.0053	0.0222	0.0033	0.0156	0.0021
$\widehat{R}_{Dcalo}$	0.0195	0.0088	0.0231	0.0053	0.0222	0.0033	0.0156	0.0021
$1/\exp(0{,}75 + b_1/8)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.848	1	0.848	1	1.1232	1	1.3591	1
$\widehat{R}_{cal}^{(1)}$	0.4107	0.3017	0.4107	0.3017	0.3705	0.0972	0.3527	0.0585
$\widehat{R}_{cal}^{(2)}$	0.0342	0.0236	0.0342	0.0236	0.0255	0.009	0.0156	0.0051
$\widehat{R}_{KL1cali}$	0.0316	0.0212	0.0316	0.0212	0.0243	0.0087	0.0149	0.005
$\widehat{R}_{KL1calr}$	0.0316	0.0212	0.0316	0.0212	0.0241	0.0087	0.0148	0.0049
$\widehat{R}_{KL1calo}$	0.0318	0.0214	0.0318	0.0214	0.0243	0.0087	0.0149	0.005
$\widehat{R}_{KL2cali}$	0.0316	0.0212	0.0316	0.0212	0.0243	0.0087	0.015	0.005
$\widehat{R}_{KL2calr}$	0.0316	0.0213	0.0316	0.0213	0.0243	0.0087	0.0149	0.005
$\widehat{R}_{KL2calo}$	0.0316	0.0213	0.0316	0.0213	0.0243	0.0087	0.0149	0.005
$\widehat{R}_{Dcali}$	0.0334	0.0231	0.0334	0.0231	0.0226	0.0085	0.014	0.0049
$\widehat{R}_{Dcalr}$	0.0332	0.023	0.0332	0.023	0.0226	0.0085	0.0139	0.0049
$\widehat{R}_{Dcalo}$	0.0332	0.023	0.0332	0.023	0.0226	0.0085	0.0139	0.0049
$1/\exp(0{,}125 + b_1/2)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	1.6643	1	1.8632	1	2.0019	1	2.2985	1
$\widehat{R}_{cal}^{(1)}$	0.2462	0.0296	0.2345	0.0194	0.228	0.0149	0.2266	0.0105
$\widehat{R}_{cal}^{(2)}$	0.0203	0.0056	0.0184	0.0033	0.0137	0.0024	0.0114	0.0014
$\widehat{R}_{KL1cali}$	0.0199	0.0053	0.0196	0.0032	0.0134	0.0023	0.0119	0.0013
$\widehat{R}_{KL1calr}$	0.0199	0.0053	0.0195	0.0032	0.0136	0.0023	0.0119	0.0013
$\widehat{R}_{KL1calo}$	0.0198	0.0052	0.0196	0.0032	0.0136	0.0023	0.012	0.0013
$\widehat{R}_{KL2cali}$	0.0198	0.0053	0.0196	0.0032	0.0134	0.0023	0.0119	0.0013
$\widehat{R}_{KL2calr}$	0.0198	0.0053	0.0196	0.0032	0.0134	0.0023	0.0119	0.0013
$\widehat{R}_{KL2calo}$	0.0198	0.0053	0.0196	0.0032	0.0134	0.0023	0.0119	0.0013
$\widehat{R}_{Dcali}$	0.0195	0.0054	0.0189	0.0033	0.0134	0.0024	0.0118	0.0013
$\widehat{R}_{Dcalr}$	0.0195	0.0054	0.0189	0.0033	0.0134	0.0024	0.0118	0.0013
$\widehat{R}_{Dcalo}$	0.0195	0.0054	0.0189	0.0033	0.0134	0.0024	0.0118	0.0013

Tabla A2.3: RB and RE for several sample sizes of the estimators of $R(0{,}9, 0{,}1)$. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY.

$1/\exp(0{,}5 + b_1/3)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.6336	1	0.6867	1	0.8431	1	0.9662	1
$\widehat{R}^{(1)}_{cal}$	1.0228	2.4065	1.0892	2.0116	1.238	1.8797	1.4022	1.9122
$\widehat{R}^{(2)}_{cal}$	0.1069	0.2128	0.0593	0.0704	0.052	0.0431	0.0357	0.0244
$\widehat{R}_{KL1cali}$	0.0908	0.1241	0.0556	0.0675	0.0476	0.0403	0.0334	0.0236
$\widehat{R}_{KL1calr}$	0.0913	0.1238	0.0552	0.0661	0.0477	0.0403	0.0334	0.0236
$\widehat{R}_{KL1calo}$	0.0913	0.1238	0.0551	0.066	0.0476	0.0403	0.0334	0.0237
$\widehat{R}_{KL2cali}$	0.0908	0.124	0.0557	0.0675	0.0477	0.0403	0.0334	0.0236
$\widehat{R}_{KL2calr}$	0.0912	0.1242	0.0558	0.0675	0.0477	0.0403	0.0335	0.0237
$\widehat{R}_{KL2calo}$	0.0912	0.1242	0.0558	0.0675	0.0477	0.0403	0.0335	0.0237
$\widehat{R}_{Dcali}$	0.0923	0.1271	0.0604	0.1281	0.0485	0.0416	0.0342	0.0244
$\widehat{R}_{Dcalr}$	0.0926	0.1273	0.0604	0.1279	0.0485	0.0416	0.0344	0.0244
$\widehat{R}_{Dcalo}$	0.0926	0.1273	0.0604	0.1279	0.0485	0.0416	0.0344	0.0244
$1/\exp(0{,}25 + b_1/2)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.8144	1	0.9677	1	1.0907	1	1.2854	1
$\widehat{R}^{(1)}_{cal}$	0.9912	1.4969	1.1483	1.4556	1.2233	1.31	1.4571	1.3506
$\widehat{R}^{(2)}_{cal}$	0.0773	0.0842	0.0534	0.0398	0.0453	0.0256	0.0315	0.0126
$\widehat{R}_{KL1cali}$	0.0704	0.075	0.0484	0.0352	0.0431	0.0236	0.0295	0.0121
$\widehat{R}_{KL1calr}$	0.07	0.0747	0.048	0.0351	0.0428	0.0236	0.0293	0.012
$\widehat{R}_{KL1calo}$	0.0708	0.0751	0.0484	0.0352	0.0432	0.0237	0.0294	0.0121
$\widehat{R}_{KL2cali}$	0.0705	0.075	0.0483	0.0352	0.0431	0.0236	0.0294	0.0121
$\widehat{R}_{KL2calr}$	0.0708	0.0753	0.0485	0.0353	0.0431	0.0236	0.0294	0.0121
$\widehat{R}_{KL2calo}$	0.0708	0.0753	0.0485	0.0353	0.0431	0.0236	0.0294	0.0121
$\widehat{R}_{Dcali}$	0.0719	0.0804	0.051	0.0369	0.0444	0.0253	0.0307	0.0125
$\widehat{R}_{Dcalr}$	0.0723	0.0806	0.0511	0.037	0.0444	0.0253	0.0306	0.0125
$\widehat{R}_{Dcalo}$	0.0723	0.0806	0.0511	0.037	0.0444	0.0253	0.0306	0.0125
$1/\exp(0{,}75 + b_1/8)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.4666	1	0.531	1	0.65	1	0.7915	1
$\widehat{R}^{(1)}_{cal}$	0.9788	2.7242	1.0596	2.6561	1.1641	2.3383	1.3844	2.5426
$\widehat{R}^{(2)}_{cal}$	0.1035	0.2375	0.0761	0.1307	0.0554	0.0734	0.0431	0.0447
$\widehat{R}_{KL1cali}$	0.0944	0.2169	0.0725	0.1268	0.0534	0.0717	0.0399	0.037
$\widehat{R}_{KL1calr}$	0.0936	0.2147	0.0722	0.1268	0.0538	0.072	0.04	0.0368
$\widehat{R}_{KL1calo}$	0.0924	0.2081	0.0726	0.1271	0.0535	0.0718	0.04	0.0371
$\widehat{R}_{KL2cali}$	0.0941	0.2168	0.0725	0.1268	0.0534	0.0717	0.0399	0.0371
$\widehat{R}_{KL2calr}$	0.0944	0.2169	0.0726	0.1269	0.0535	0.072	0.04	0.0371
$\widehat{R}_{KL2calo}$	0.0944	0.2169	0.0726	0.1269	0.0535	0.072	0.04	0.0371
$\widehat{R}_{Dcali}$	0.0989	0.2464	0.0728	0.1289	0.0551	0.0737	0.0408	0.0379
$\widehat{R}_{Dcalr}$	0.099	0.2465	0.0728	0.1289	0.0551	0.0737	0.041	0.038
$\widehat{R}_{Dcalo}$	0.099	0.2465	0.0728	0.1289	0.0551	0.0737	0.041	0.038
$1/\exp(0{,}125 + b_1/2)$								
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	1.0535	1	1.1436	1	1.2894	1	1.5446	1
$\widehat{R}^{(1)}_{cal}$	0.9491	0.9409	1.0499	1.0116	1.1124	0.9087	1.3752	0.9261
$\widehat{R}^{(2)}_{cal}$	0.0468	0.0363	0.0508	0.0281	0.038	0.0159	0.0364	0.0087
$\widehat{R}_{KL1cali}$	0.0401	0.0319	0.0463	0.0254	0.0386	0.0157	0.0361	0.0082
$\widehat{R}_{KL1calr}$	0.0403	0.032	0.0463	0.0252	0.0385	0.0156	0.0362	0.0082
$\widehat{R}_{KL1calo}$	0.0397	0.0317	0.0465	0.0253	0.0385	0.0156	0.0362	0.0082
$\widehat{R}_{KL2cali}$	0.0402	0.032	0.0463	0.0254	0.0386	0.0157	0.0361	0.0082
$\widehat{R}_{KL2calr}$	0.0401	0.032	0.0464	0.0254	0.0386	0.0157	0.0361	0.0082
$\widehat{R}_{KL2calo}$	0.0401	0.032	0.0464	0.0254	0.0386	0.0157	0.0361	0.0082
$\widehat{R}_{Dcali}$	0.0434	0.0347	0.0481	0.0274	0.0396	0.0164	0.0362	0.0083
$\widehat{R}_{Dcalr}$	0.0434	0.0348	0.0482	0.0274	0.0396	0.0164	0.0362	0.0083
$\widehat{R}_{Dcalo}$	0.0434	0.0348	0.0482	0.0274	0.0396	0.0164	0.0362	0.0083

Tabla A2.4: RB and RE for several sample sizes of the estimators of $R(0{,}9, 0{,}2)$. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY.

				$1/\exp(0{,}5 + b_1/3)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.5727	1	0.683	1	0.7759	1	0.9313	1
$\widehat{R}^{(1)}_{cal}$	0.9516	1.9501	1.0774	1.987	1.2049	2.0104	1.4164	2.0432
$\widehat{R}^{(2)}_{cal}$	0.0293	0.0526	0.0277	0.034	0.0225	0.0214	0.0157	0.0117
$\widehat{R}_{KL1cali}$	0.0283	0.0452	0.027	0.0309	0.0212	0.0208	0.0151	0.0115
$\widehat{R}_{KL1calr}$	0.029	0.0451	0.0269	0.0309	0.0214	0.0209	0.0152	0.0115
$\widehat{R}_{KL1calo}$	0.0291	0.0453	0.0271	0.0309	0.0212	0.0207	0.0152	0.0115
$\widehat{R}_{KL2cali}$	0.0286	0.0453	0.0271	0.0308	0.0212	0.0208	0.0151	0.0115
$\widehat{R}_{KL2calr}$	0.0288	0.0454	0.0272	0.0309	0.0212	0.0208	0.0152	0.0115
$\widehat{R}_{KL2calo}$	0.0288	0.0454	0.0272	0.0309	0.0212	0.0208	0.0152	0.0115
$\widehat{R}_{Dcali}$	0.0309	0.0471	0.0268	0.0315	0.0222	0.0209	0.0156	0.0117
$\widehat{R}_{Dcalr}$	0.0311	0.0472	0.027	0.0316	0.0222	0.0209	0.0157	0.0117
$\widehat{R}_{Dcalo}$	0.0311	0.0472	0.027	0.0316	0.0222	0.0209	0.0157	0.0117
				$1/\exp(0{,}25 + b_1/2)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.7775	1	0.8934	1	1.0167	1	1.2284	1
$\widehat{R}^{(1)}_{cal}$	0.9419	1.3871	1.0599	1.3864	1.2055	1.3996	1.4514	1.4305
$\widehat{R}^{(2)}_{cal}$	0.0105	0.0219	0.019	0.0189	0.0193	0.0122	0.014	0.0066
$\widehat{R}_{KL1cali}$	0.0067	0.0203	0.0184	0.0182	0.0191	0.0119	0.0136	0.0065
$\widehat{R}_{KL1calr}$	0.0069	0.0203	0.0181	0.0181	0.0191	0.0119	0.0135	0.0064
$\widehat{R}_{KL1calo}$	0.0067	0.0203	0.0182	0.0181	0.0191	0.0119	0.0136	0.0064
$\widehat{R}_{KL2cali}$	0.0066	0.0203	0.0184	0.0182	0.0191	0.0119	0.0136	0.0065
$\widehat{R}_{KL2calr}$	0.0067	0.0203	0.0183	0.0182	0.0191	0.0119	0.0136	0.0065
$\widehat{R}_{KL2calo}$	0.0067	0.0203	0.0183	0.0182	0.0191	0.0119	0.0136	0.0065
$\widehat{R}_{Dcali}$	0.0074	0.0219	0.0207	0.0193	0.0199	0.0124	0.014	0.0066
$\widehat{R}_{Dcalr}$	0.0074	0.0219	0.0207	0.0193	0.0199	0.0124	0.014	0.0066
$\widehat{R}_{Dcalo}$	0.0074	0.0219	0.0207	0.0193	0.0199	0.0124	0.014	0.0066
				$1/\exp(0{,}75 + b_1/8)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.4117	1	0.4871	1	0.5832	1	0.7334	1
$\widehat{R}^{(2)}_{cal}$	0.8896	2.348	0.9978	2.4013	1.1243	2.4448	1.3552	2.5469
$\widehat{R}^{(2)}_{cal}$	0.0387	0.0783	0.0272	0.0491	0.0235	0.0338	0.0147	0.0178
$\widehat{R}_{KL1cali}$	0.035	0.0713	0.024	0.0459	0.0224	0.0322	0.0148	0.0174
$\widehat{R}_{KL1calr}$	0.0365	0.0719	0.0237	0.0458	0.0225	0.0323	0.0148	0.0173
$\widehat{R}_{KL1calo}$	0.0356	0.0715	0.024	0.0459	0.0223	0.0322	0.0148	0.0173
$\widehat{R}_{KL2cali}$	0.0353	0.0713	0.024	0.0459	0.0224	0.0322	0.0148	0.0174
$\widehat{R}_{KL2calr}$	0.0357	0.0716	0.0241	0.0459	0.0224	0.0322	0.0149	0.0174
$\widehat{R}_{KL2calo}$	0.0357	0.0716	0.0241	0.0459	0.0224	0.0322	0.0149	0.0174
$\widehat{R}_{Dcali}$	0.0359	0.0744	0.0251	0.0478	0.0231	0.0329	0.0145	0.0174
$\widehat{R}_{Dcalr}$	0.0366	0.0748	0.0251	0.0478	0.0231	0.0329	0.0145	0.0174
$\widehat{R}_{Dcalo}$	0.0366	0.0748	0.0251	0.0478	0.0231	0.0329	0.0145	0.0174
				$1/\exp(0{,}125 + b_1/2)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	1.0141	1	1.119	1	1.2479	1	1.4519	1
$\widehat{R}^{(1)}_{cal}$	0.9229	0.9531	0.9997	0.9341	1.1202	0.9487	1.2918	0.9429
$\widehat{R}^{(2)}_{cal}$	0.0221	0.0158	0.016	0.0109	0.0149	0.0077	0.014	0.0045
$\widehat{R}_{KL1cali}$	0.0222	0.0156	0.0154	0.0104	0.0146	0.0076	0.0139	0.0044
$\widehat{R}_{KL1calr}$	0.0228	0.0156	0.0155	0.0104	0.0145	0.0076	0.0138	0.0044
$\widehat{R}_{KL1calo}$	0.0227	0.0156	0.0155	0.0104	0.0146	0.0076	0.0138	0.0044
$\widehat{R}_{KL2cali}$	0.0223	0.0156	0.0154	0.0104	0.0146	0.0076	0.0139	0.0044
$\widehat{R}_{KL2calr}$	0.0225	0.0156	0.0155	0.0104	0.0146	0.0076	0.014	0.0044
$\widehat{R}_{KL2calo}$	0.0225	0.0156	0.0155	0.0104	0.0146	0.0076	0.014	0.0044
$\widehat{R}_{Dcali}$	0.0237	0.0165	0.0163	0.0108	0.0153	0.0078	0.0146	0.0045
$\widehat{R}_{Dcalr}$	0.0239	0.0165	0.0163	0.0108	0.0153	0.0078	0.0147	0.0045
$\widehat{R}_{Dcalo}$	0.0239	0.0165	0.0163	0.0108	0.0153	0.0078	0.0147	0.0045

Tabla A2.5: RB and RE for several sample sizes of the estimators of $R(0{,}95, 0{,}2)$. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY.

				$1/\exp(0{,}5 + b_1/3)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.2525	1	0.3418	1	0.4087	1	0.5525	1
$\widehat{R}^{(1)}_{cal}$	0.6794	2.4489	0.8063	2.5222	0.8924	2.5047	1.0903	2.5143
$\widehat{R}^{(2)}_{cal}$	0.0519	0.1743	0.0374	0.0959	0.033	0.0638	0.023	0.0322
$\widehat{R}_{KL1cali}$	0.0418	0.1339	0.0339	0.0882	0.0312	0.0602	0.0222	0.0311
$\widehat{R}_{KL1calr}$	0.0414	0.1327	0.0343	0.0883	0.0311	0.0603	0.0219	0.0309
$\widehat{R}_{KL1calo}$	0.042	0.1334	0.0342	0.0882	0.0311	0.0602	0.0221	0.031
$\widehat{R}_{KL2cali}$	0.0418	0.1339	0.0338	0.0882	0.0312	0.0602	0.0222	0.0311
$\widehat{R}_{KL2calr}$	0.0421	0.134	0.034	0.0882	0.0312	0.0602	0.0223	0.0311
$\widehat{R}_{KL2calo}$	0.0421	0.134	0.034	0.0882	0.0312	0.0602	0.0223	0.0311
$\widehat{R}_{Dcali}$	0.0444	0.145	0.0368	0.0946	0.0346	0.0649	0.023	0.0325
$\widehat{R}_{Dcalr}$	0.0446	0.145	0.0369	0.0947	0.0346	0.0649	0.0231	0.0325
$\widehat{R}_{Dcalo}$	0.0446	0.145	0.0369	0.0947	0.0346	0.0649	0.0231	0.0325
				$1/\exp(0{,}25 + b_1/2)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.4557	1	0.5132	1	0.5897	1	0.7267	1
$\widehat{R}^{(1)}_{cal}$	0.773	1.8254	0.8512	1.8864	0.9553	1.9073	1.1336	1.9188
$\widehat{R}^{(2)}_{cal}$	0.048	0.0928	0.0317	0.0538	0.0266	0.0411	0.0215	0.02
$\widehat{R}_{KL1cali}$	0.0443	0.0702	0.0297	0.0511	0.0232	0.0366	0.0231	0.0199
$\widehat{R}_{KL1calr}$	0.0441	0.0702	0.0293	0.0502	0.0232	0.0366	0.0229	0.02
$\widehat{R}_{KL1calo}$	0.0441	0.0697	0.0295	0.0503	0.0235	0.0367	0.0229	0.0199
$\widehat{R}_{KL2cali}$	0.0443	0.0702	0.0297	0.0512	0.0232	0.0366	0.0231	0.02
$\widehat{R}_{KL2calr}$	0.0446	0.0704	0.0298	0.0512	0.0232	0.0366	0.0231	0.02
$\widehat{R}_{KL2calo}$	0.0446	0.0704	0.0298	0.0512	0.0232	0.0366	0.0231	0.02
$\widehat{R}_{Dcali}$	0.0484	0.0829	0.0322	0.054	0.0247	0.038	0.0235	0.0201
$\widehat{R}_{Dcalr}$	0.0487	0.083	0.0323	0.0541	0.0247	0.038	0.0235	0.0201
$\widehat{R}_{Dcalo}$	0.0487	0.083	0.0323	0.0541	0.0247	0.038	0.0235	0.0201
				$1/\exp(0{,}75 + b_1/8)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.1243	1	0.1843	1	0.2609	1	0.389	1
$\widehat{R}^{(1)}_{cal}$	0.6477	3.2315	0.7304	3.1825	0.8417	3.2802	1.0211	3.2011
$\widehat{R}^{(2)}_{cal}$	0.0583	0.2974	0.0409	0.1495	0.0359	0.1049	0.0251	0.0512
$\widehat{R}_{KL1cali}$	0.0455	0.228	0.0414	0.1439	0.0311	0.092	0.0225	0.049
$\widehat{R}_{KL1calr}$	0.0464	0.2295	0.0415	0.1444	0.0308	0.0917	0.0227	0.0491
$\widehat{R}_{KL1calo}$	0.0462	0.2279	0.0416	0.1439	0.031	0.0921	0.0227	0.0491
$\widehat{R}_{KL2cali}$	0.0457	0.2278	0.0415	0.144	0.0311	0.092	0.0224	0.049
$\widehat{R}_{KL2calr}$	0.0462	0.2285	0.0416	0.144	0.0311	0.092	0.0226	0.0491
$\widehat{R}_{KL2calo}$	0.0462	0.2285	0.0416	0.144	0.0311	0.092	0.0226	0.0491
$\widehat{R}_{Dcali}$	0.0451	0.2283	0.0459	0.1554	0.0337	0.0975	0.0238	0.0507
$\widehat{R}_{Dcalr}$	0.046	0.2289	0.0459	0.1554	0.0337	0.0975	0.024	0.0507
$\widehat{R}_{Dcalo}$	0.046	0.2289	0.0459	0.1554	0.0337	0.0975	0.024	0.0507
				$1/\exp(0{,}125 + b_1/2)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.587	1	0.703	1	0.8148	1	0.9631	1
$\widehat{R}^{(1)}_{cal}$	0.7991	1.488	0.945	1.5212	1.0747	1.51	1.2532	1.5233
$\widehat{R}^{(2)}_{cal}$	0.039	0.0577	0.0293	0.0352	0.0272	0.0218	0.0192	0.0122
$\widehat{R}_{KL1cali}$	0.0312	0.0483	0.0277	0.0319	0.0265	0.0212	0.0189	0.012
$\widehat{R}_{KL1calr}$	0.0318	0.0486	0.0282	0.0321	0.0262	0.0211	0.019	0.012
$\widehat{R}_{KL1calo}$	0.0315	0.0485	0.0279	0.032	0.0262	0.0211	0.0191	0.012
$\widehat{R}_{KL2cali}$	0.0311	0.0483	0.0277	0.0319	0.0264	0.0212	0.0189	0.012
$\widehat{R}_{KL2calr}$	0.0313	0.0484	0.0278	0.0319	0.0265	0.0212	0.019	0.012
$\widehat{R}_{KL2calo}$	0.0313	0.0484	0.0278	0.0319	0.0265	0.0212	0.019	0.012
$\widehat{R}_{Dcali}$	0.0304	0.0496	0.0295	0.0329	0.0272	0.0218	0.0198	0.0123
$\widehat{R}_{Dcalr}$	0.0306	0.0497	0.0294	0.0329	0.0272	0.0218	0.0198	0.0124
$\widehat{R}_{Dcalo}$	0.0306	0.0497	0.0294	0.033	0.0272	0.0218	0.0198	0.0124

Tabla A2.6: RB and RE for several sample sizes of the estimators of $R(0{,}95, 0{,}5)$. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY.

				$1/\exp(0{,}5 + b_1/3)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	-0.4625	1	-0.4695	1	-0.4355	1	-0.4268	1
$\widehat{R}^{(1)}_{cal}$	0.5605	2.9425	0.6607	3.5969	0.7903	4.6128	0.9173	5.4401
$\widehat{R}^{(2)}_{cal}$	0.0305	0.2341	0.0106	0.1129	0.0129	0.1031	0.0115	0.0672
$\widehat{R}_{KL1cali}$	0.0221	0.1788	0.0101	0.1072	0.0113	0.0919	0.0109	0.0655
$\widehat{R}_{KL1calr}$	0.0223	0.1756	0.0109	0.1079	0.0114	0.0914	0.0114	0.0654
$\widehat{R}_{KL1calo}$	0.0229	0.1786	0.0108	0.1076	0.0114	0.0918	0.0114	0.0656
$\widehat{R}_{KL2cali}$	0.022	0.1786	0.0102	0.1075	0.0113	0.0919	0.0109	0.0655
$\widehat{R}_{KL2calr}$	0.0232	0.1788	0.0107	0.1075	0.0114	0.0919	0.0115	0.0656
$\widehat{R}_{KL2calo}$	0.0232	0.1788	0.0107	0.1075	0.0114	0.0919	0.0115	0.0656
$\widehat{R}_{Dcali}$	0.0247	0.1849	0.0124	0.1181	0.0142	0.0971	0.0109	0.0665
$\widehat{R}_{Dcalr}$	0.0258	0.1848	0.0129	0.1183	0.0142	0.0971	0.0115	0.0666
$\widehat{R}_{Dcalo}$	0.0258	0.1848	0.0129	0.1183	0.0142	0.0971	0.0115	0.0666
				$1/\exp(0{,}25 + b_1/2)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	-0.0577	1	0.0351	1	0.1056	1	0.2102	1
$\widehat{R}^{(1)}_{cal}$	0.6081	3.736	0.8029	4.5287	0.8816	4.7993	1.0674	5.0535
$\widehat{R}^{2}_{cal}$	0.0235	0.208	0.0201	0.1103	0.0111	0.0791	0.0104	0.0421
$\widehat{R}_{KL1cali}$	0.0195	0.1675	0.0211	0.1072	0.0103	0.0777	0.0093	0.0413
$\widehat{R}_{KL1calr}$	0.0207	0.1683	0.0218	0.1076	0.0101	0.0759	0.0094	0.0413
$\widehat{R}_{KL1calo}$	0.0205	0.1679	0.0215	0.1072	0.0101	0.076	0.0094	0.0412
$\widehat{R}_{KL2cali}$	0.0195	0.1674	0.0212	0.1073	0.0103	0.0776	0.0093	0.0413
$\widehat{R}_{KL2calr}$	0.0204	0.168	0.0219	0.1074	0.0104	0.0776	0.0096	0.0414
$\widehat{R}_{KL2calo}$	0.0204	0.168	0.0219	0.1074	0.0104	0.0776	0.0096	0.0414
$\widehat{R}_{Dcali}$	0.0205	0.1768	0.0224	0.1137	0.0131	0.0809	0.0104	0.043
$\widehat{R}_{Dcalr}$	0.0217	0.1774	0.023	0.114	0.0131	0.0809	0.0107	0.0431
$\widehat{R}_{Dcalo}$	0.0217	0.1774	0.023	0.114	0.0131	0.0809	0.0107	0.0431
				$1/\exp(0{,}75 + b_1/8)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	-0.6162	1	-0.6236	1	-0.6225	1	-0.6319	1
$\widehat{R}_{cal1}$	0.5322	2.6084	0.6176	2.8892	0.7233	3.314	0.8914	4.2668
$\widehat{R}^{2}_{cal}$	0.027	0.2134	0.017	0.1204	0.0163	0.0954	0.0092	0.059
$\widehat{R}_{KL1cali}$	0.0126	0.1638	0.0187	0.113	0.0109	0.0818	0.007	0.0576
$\widehat{R}_{KL1calr}$	0.0138	0.1636	0.0189	0.113	0.0108	0.0812	0.0077	0.0576
$\widehat{R}_{KL1calo}$	0.014	0.1645	0.019	0.1131	0.011	0.0818	0.0077	0.0577
$\widehat{R}_{KL2cali}$	0.0125	0.1635	0.0187	0.1129	0.011	0.0818	0.007	0.0576
$\widehat{R}_{KL2calr}$	0.014	0.1644	0.0191	0.1131	0.011	0.0818	0.0077	0.0577
$\widehat{R}_{KL2calo}$	0.014	0.1644	0.0191	0.1131	0.011	0.0818	0.0077	0.0577
$\widehat{R}_{Dcali}$	0.015	0.1697	0.0231	0.1266	0.0145	0.0886	0.0087	0.0594
$\widehat{R}_{Dcalr}$	0.0165	0.1701	0.0234	0.1267	0.0145	0.0886	0.0094	0.0595
$\widehat{R}_{Dcalo}$	0.0165	0.1701	0.0234	0.1267	0.0145	0.0886	0.0094	0.0595
				$1/\exp(0{,}125 + b_1/2)$				
	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 100$		$n = 125$		$n = 150$		$n = 200$	
$\widehat{R}_{HT}$	0.2315	1	0.3475	1	0.423	1	0.5538	1
$\widehat{R}_{cal1}$	0.6754	2.5757	0.8457	2.5934	0.9563	2.7198	1.1544	2.8213
$\widehat{R}^{2}_{cal}$	0.0171	0.1007	0.0112	0.0464	0.011	0.0319	0.0041	0.0181
$\widehat{R}_{KL1cali}$	0.0125	0.0877	0.013	0.046	0.0109	0.0308	0.0025	0.0178
$\widehat{R}_{KL1calr}$	0.0125	0.0871	0.0128	0.0458	0.0108	0.0305	0.0028	0.0178
$\widehat{R}_{KL1calo}$	0.0128	0.0872	0.013	0.0459	0.0111	0.0309	0.0028	0.0178
$\widehat{R}_{KL2cali}$	0.0125	0.0878	0.013	0.0459	0.011	0.0308	0.0025	0.0178
$\widehat{R}_{KL2calr}$	0.0133	0.0881	0.0131	0.046	0.011	0.0308	0.0028	0.0178
$\widehat{R}_{KL2calo}$	0.0133	0.0881	0.0131	0.046	0.011	0.0308	0.0028	0.0178
$\widehat{R}_{Dcali}$	0.0156	0.0983	0.0147	0.0477	0.0123	0.032	0.0029	0.0181
$\widehat{R}_{Dcalr}$	0.0165	0.0985	0.0148	0.0477	0.0123	0.032	0.0032	0.0181
$\widehat{R}_{Dcalo}$	0.0165	0.0985	0.0148	0.0477	0.0123	0.032	0.0032	0.0181

Tabla A2.7: AL, CP %,L % and U % for several sample sizes and several resampling method of the estimators compared. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY $R(0,9,0,1)$

.

	AL	CP %	L %	U %	AL	CP %	L %	U %	AL	CP %	L %	U %	AL	CP %	L %	U %
									Booth et al., (1994)							
Estimator	n = 100				n = 125				n = 150				n = 200			
$\widehat{R}_{HT}$	9.4346	100	0	0	7.85	98.1982	1.8018	0	6.0947	96.7033	3.2967	0	4.7178	96.5517	3.4483	0
$\widehat{R}^1_{cal}$	7.2357	100	0	0	5.8027	98.1982	0.9009	0.9009	5.1583	96.7033	2.7473	0.5495	3.7372	96.5517	2.5862	0.8621
$\widehat{R}^2_{cal}$	5.0012	99.0476	0	0.9524	3.8825	95.4955	0	4.5045	3.8122	97.8022	0	2.1978	2.9728	94.8276	0.8621	4.3103
$\widehat{R}_{KL1cali}$	5.1028	99.0476	0	0.9524	3.9338	95.4955	0	4.5045	3.7918	97.8022	0.5495	1.6484	2.9855	96.5517	0.8621	2.5862
$\widehat{R}_{KL1calr}$	5.1026	99.0476	0	0.9524	3.9327	95.4955	0	4.5045	3.7871	97.8022	0.5495	1.6484	2.9854	95.6897	0.8621	3.4483
$\widehat{R}_{KL1calo}$	5.1006	99.0476	0	0.9524	3.9338	95.4955	0	4.5045	3.7887	97.8022	0.5495	1.6484	2.9854	96.5517	0.8621	2.5862
$\widehat{R}_{KL2cali}$	5.1028	99.0476	0	0.9524	3.9337	95.4955	0	4.5045	3.7918	97.8022	0.5495	1.6484	2.9851	96.5517	0.8621	2.5862
$\widehat{R}_{KL2calr}$	5.1025	99.0476	0	0.9524	3.9337	95.4955	0	4.5045	3.7918	97.8022	0.5495	1.6484	2.9848	96.5517	0.8621	2.5862
$\widehat{R}_{KL2calo}$	5.1025	99.0476	0	0.9524	3.9337	95.4955	0	4.5045	3.7918	97.8022	0.5495	1.6484	2.9848	96.5517	0.8621	2.5862
$\widehat{R}_{Dcali}$	5.1468	99.0476	0	0.9524	3.9553	96.3964	0	3.6036	3.8528	97.2527	0.5495	2.1978	3.0008	96.5517	0.8621	2.5862
$\widehat{R}_{Dcalr}$	5.1465	99.0476	0	0.9524	3.955	96.3964	0	3.6036	3.8529	97.2527	0.5495	2.1978	3.0004	96.5517	0.8621	2.5862
$\widehat{R}_{Dcalo}$	5.1465	99.0476	0	0.9524	3.955	96.3964	0	3.6036	3.8529	97.2527	0.5495	2.1978	3.0004	96.5517	0.8621	2.5862
									Antal, E., Tillé, Y. (2014)							
$\widehat{R}_{HT}$	7.3592	100	0	0	5.5794	95.4955	3.6036	0.9009	5.7036	96.7033	2.7473	0.5495	4.1913	95.6897	4.3103	0
$\widehat{R}^1_{cal}$	8.5844	99.0476	0	0.9524	6.4928	96.3964	0	3.6036	6.5833	97.2527	1.0989	1.6484	4.6257	97.4138	0.8621	1.7241
$\widehat{R}^2_{cal}$	5.1692	98.0952	0	1.9048	3.9134	94.5946	0	5.4054	3.9682	96.7033	0.5495	2.7473	3.07	94.8276	1.7241	3.4483
$\widehat{R}_{KL1cali}$	4.9174	98.0952	0	1.9048	3.8008	94.5946	0.9009	4.5045	3.8267	96.7033	1.0989	2.1978	3.0318	95.6897	1.7241	2.5862
$\widehat{R}_{KL1calr}$	4.8834	98.0952	0	1.9048	3.793	94.5946	0.9009	4.5045	3.8258	96.7033	1.0989	2.1978	3.0298	95.6897	1.7241	2.5862
$\widehat{R}_{KL1calo}$	4.9093	98.0952	0	1.9048	3.7945	94.5946	0.9009	4.5045	3.8258	96.7033	1.0989	2.1978	3.03	95.6897	1.7241	2.5862
$\widehat{R}_{KL2cali}$	4.9174	98.0952	0	1.9048	3.8018	94.5946	0.9009	4.5045	3.8267	96.7033	1.0989	2.1978	3.0318	95.6897	1.7241	2.5862
$\widehat{R}_{KL2calr}$	4.9178	98.0952	0	1.9048	3.8019	94.5946	0.9009	4.5045	3.8272	96.7033	1.0989	2.1978	3.032	95.6897	1.7241	2.5862
$\widehat{R}_{KL2calo}$	4.9181	98.0952	0	1.9048	3.8018	94.5946	0.9009	4.5045	3.8273	96.7033	1.0989	2.1978	3.032	95.6897	1.7241	2.5862
$\widehat{R}_{Dcali}$	4.9548	98.0952	0	1.9048	3.8577	93.6937	0.9009	5.4054	3.8899	97.2527	0.5495	2.1978	3.066	96.5517	1.7241	1.7241
$\widehat{R}_{Dcalr}$	4.9548	98.0952	0	1.9048	3.859	93.6937	0.9009	5.4054	3.8901	97.2527	0.5495	2.1978	3.0661	96.5517	1.7241	1.7241
$\widehat{R}_{Dcalo}$	4.9538	98.0952	0	1.9048	3.859	93.6937	0.9009	5.4054	3.89	97.2527	0.5495	2.1978	3.0661	96.5517	1.7241	1.7241
									Antal, E., Tillé, Y. (2011).							
$\widehat{R}_{HT}$	7.3742	99.0476	0	0.9524	6.4673	99.0991	0	0.9009	6.0678	96.7033	3.2967	0	4.3967	95.6897	4.3103	0
$\widehat{R}^1_{cal}$	8.8376	99.0476	0	0.9524	6.9812	94.5946	0	5.4054	6.5299	96.7033	1.0989	2.1978	4.3094	98.2759	0.8621	0.8621
$\widehat{R}^2_{cal}$	5.1544	98.0952	0	1.9048	4.0739	93.6937	0	6.3063	3.9248	96.1538	1.0989	2.7473	3.0807	96.5517	0.8621	2.5862
$\widehat{R}_{KL1cali}$	5.0125	98.0952	0	1.9048	3.932	93.6937	0.9009	5.4054	3.9004	95.6044	1.0989	3.2967	3.0634	96.5517	0.8621	2.5862
$\widehat{R}_{KL1calr}$	4.9587	98.0952	0	1.9048	3.9485	93.6937	0.9009	5.4054	3.8933	95.6044	1.0989	3.2967	3.0686	96.5517	0.8621	2.5862
$\widehat{R}_{KL1calo}$	4.9565	98.0952	0	1.9048	3.9485	93.6937	0.9009	5.4054	3.8942	95.6044	1.0989	3.2967	3.0676	96.5517	0.8621	2.5862
$\widehat{R}_{KL2cali}$	5.0126	98.0952	0	1.9048	3.9319	93.6937	0.9009	5.4054	3.9004	95.6044	1.0989	3.2967	3.0633	96.5517	0.8621	2.5862
$\widehat{R}_{KL2calr}$	5.0116	98.0952	0	1.9048	3.9322	93.6937	0.9009	5.4054	3.9006	95.6044	1.0989	3.2967	3.0636	96.5517	0.8621	2.5862
$\widehat{R}_{KL2calo}$	5.0116	98.0952	0	1.9048	3.9322	93.6937	0.9009	5.4054	3.9006	95.6044	1.0989	3.2967	3.0636	96.5517	0.8621	2.5862
$\widehat{R}_{Dcali}$	5.1284	98.0952	0	1.9048	4.0927	94.5946	0.9009	4.5045	3.9496	96.7033	0.5495	2.7473	3.0956	97.4138	0.8621	1.7241
$\widehat{R}_{Dcalr}$	5.1269	98.0952	0	1.9048	4.0928	94.5946	0.9009	4.5045	3.9497	96.7033	0.5495	2.7473	3.0958	97.4138	0.8621	1.7241
$\widehat{R}_{Dcalo}$	5.1269	98.0952	0	1.9048	4.0928	94.5946	0.9009	4.5045	3.9497	96.7033	0.5495	2.7473	3.0958	97.4138	0.8621	1.7241

Tabla A2.8: AL, CP %,L % and U % for several sample sizes and several resampling method of the estimators compared. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY $R(0,9,0,2)$

.

	AL	CP %	L %	U %	AL	CP %	L %	U %	AL	CP %	L %	U %	AL	CP %	L %	U %
									Booth et al., (1994)							
Estimator	$n = 100$				$n = 125$				$n = 150$				$n = 200$			
$\widehat{R}_{HT}$	4.9792	100	0	0	4.8907	98.6014	1.3986	0	3.4594	97.5	2.5	0	2.6905	94.359	5.641	0
$\widehat{R}^1_{cal}$	4.7952	97.2973	2.7027	0	3.8111	96.5035	1.3986	2.0979	2.8273	97.5	1.5	1	2.194	96.4103	2.5641	1.0256
$\widehat{R}^2_{cal}$	2.7512	98.1982	1.8018	0	2.276	96.5035	0.6993	2.7972	1.8653	93.5	3	3.5	1.5002	95.3846	0.5128	4.1026
$\widehat{R}_{KL1cali}$	2.8209	98.1982	1.8018	0	2.1877	95.8042	0	4.1958	1.8608	94	2	4	1.5071	94.8718	0.5128	4.6154
$\widehat{R}_{KL1calr}$	2.8187	98.1982	1.8018	0	2.1875	95.8042	0	4.1958	1.8584	94	2	4	1.5076	94.8718	0.5128	4.6154
$\widehat{R}_{KL1calo}$	2.8184	98.1982	1.8018	0	2.1869	95.8042	0	4.1958	1.8585	94	2	4	1.5069	94.8718	0.5128	4.6154
$\widehat{R}_{KL2cali}$	2.8208	98.1982	1.8018	0	2.1878	95.8042	0	4.1958	1.8607	94	2	4	1.5071	94.8718	0.5128	4.6154
$\widehat{R}_{KL2calr}$	2.8218	98.1982	1.8018	0	2.1876	95.8042	0	4.1958	1.8613	94	2	4	1.507	94.8718	0.5128	4.6154
$\widehat{R}_{KL2calo}$	2.8203	98.1982	1.8018	0	2.1877	95.8042	0	4.1958	1.8613	94	2	4	1.507	94.8718	0.5128	4.6154
$\widehat{R}_{Dcali}$	2.7554	98.1982	1.8018	0	2.1969	95.1049	0.6993	4.1958	1.8771	94	2	4	1.5225	95.3846	0.5128	4.1026
$\widehat{R}_{Dcalr}$	2.756	98.1982	1.8018	0	2.1967	95.1049	0.6993	4.1958	1.8775	94	2	4	1.5223	95.3846	0.5128	4.1026
$\widehat{R}_{Dcalo}$	2.756	98.1982	1.8018	0	2.1967	95.1049	0.6993	4.1958	1.8769	94	2	4	1.5223	95.3846	0.5128	4.1026
									Antal, E., Tillé, Y. (2014)							
$\widehat{R}_{HT}$	3.2056	99.0991	0.9009	0	3.8494	97.2028	2.7972	0	2.8248	94	5.5	0.5	2.2699	93.8462	6.1538	0
$\widehat{R}^1_{cal}$	5.125	100	0	0	4.3179	95.8042	0	4.1958	3.4014	97	0	3	2.5385	98.4615	0	1.5385
$\widehat{R}^2_{cal}$	2.3617	100	0	0	1.9477	95.1049	1.3986	3.4965	1.7589	95.5	1.5	3	1.4667	93.8462	1.5385	4.6154
$\widehat{R}_{KL1cali}$	2.3224	100	0	0	1.9342	95.8042	0.6993	3.4965	1.7262	94	1.5	4.5	1.4566	94.359	1.0256	4.6154
$\widehat{R}_{KL1calr}$	2.3212	100	0	0	1.9308	95.8042	0.6993	3.4965	1.7262	94	1.5	4.5	1.4565	94.359	1.0256	4.6154
$\widehat{R}_{KL1calo}$	2.3219	100	0	0	1.9312	95.8042	0.6993	3.4965	1.726	94	1.5	4.5	1.4567	94.359	1.0256	4.6154
$\widehat{R}_{KL2cali}$	2.322	100	0	0	1.934	95.8042	0.6993	3.4965	1.7263	94	1.5	4.5	1.4566	94.359	1.0256	4.6154
$\widehat{R}_{KL2calr}$	2.3215	100	0	0	1.9342	95.8042	0.6993	3.4965	1.7266	94	1.5	4.5	1.4566	94.359	1.0256	4.6154
$\widehat{R}_{KL2calo}$	2.3232	100	0	0	1.9341	95.8042	0.6993	3.4965	1.7266	94	1.5	4.5	1.4566	94.359	1.0256	4.6154
$\widehat{R}_{Dcali}$	2.3315	100	0	0	1.9623	95.8042	0.6993	3.4965	1.7419	93.5	1.5	5	1.4712	94.8718	0.5128	4.6154
$\widehat{R}_{Dcalr}$	2.3312	100	0	0	1.9623	95.8042	0.6993	3.4965	1.7422	93.5	1.5	5	1.4712	94.8718	0.5128	4.6154
$\widehat{R}_{Dcalo}$	2.3313	100	0	0	1.9622	95.8042	0.6993	3.4965	1.7423	93.5	1.5	5	1.4712	94.8718	0.5128	4.6154
									Antal, E., Tillé, Y. (2011).							
$\widehat{R}_{HT}$	3.6236	99.0991	0.9009	0	4.1502	97.2028	2.7972	0	2.9825	96	3.5	0.5	2.4595	94.8718	5.1282	0
$\widehat{R}^1_{cal}$	5.2815	100	0	0	4.247	96.5035	0	3.4965	3.4858	96.5	0	3.5	2.3496	96.9231	1.0256	2.0513
$\widehat{R}^2_{cal}$	2.3619	100	0	0	1.9309	95.8042	0.6993	3.4965	1.7541	94.5	1.5	4	1.4434	94.359	1.5385	4.1026
$\widehat{R}_{KL1cali}$	2.4016	100	0	0	1.952	95.1049	0.6993	4.1958	1.7537	93.5	3	3.5	1.4483	94.8718	1.0256	4.1026
$\widehat{R}_{KL1calr}$	2.3967	100	0	0	1.9501	95.1049	0.6993	4.1958	1.7535	93.5	3	3.5	1.4479	95.3846	0.5128	4.1026
$\widehat{R}_{KL1calo}$	2.3988	100	0	0	1.95	95.1049	0.6993	4.1958	1.753	93.5	3	3.5	1.4479	95.3846	0.5128	4.1026
$\widehat{R}_{KL2cali}$	2.4016	100	0	0	1.9519	95.1049	0.6993	4.1958	1.7536	93.5	3	3.5	1.4483	94.8718	1.0256	4.1026
$\widehat{R}_{KL2calr}$	2.4011	100	0	0	1.9521	95.1049	0.6993	4.1958	1.7537	93.5	3	3.5	1.4483	94.8718	1.0256	4.1026
$\widehat{R}_{KL2calo}$	2.4012	100	0	0	1.952	95.1049	0.6993	4.1958	1.7537	93.5	3	3.5	1.4483	94.8718	1.0256	4.1026
$\widehat{R}_{Dcali}$	2.4351	100	0	0	1.9803	95.8042	0.6993	3.4965	1.7689	93.5	3	3.5	1.4618	94.359	1.5385	4.1026
$\widehat{R}_{Dcalr}$	2.435	100	0	0	1.9802	95.8042	0.6993	3.4965	1.7689	93.5	3	3.5	1.4617	94.359	1.5385	4.1026
$\widehat{R}_{Dcalo}$	2.435	100	0	0	1.9802	95.8042	0.6993	3.4965	1.7689	93.5	3	3.5	1.4617	94.359	1.5385	4.1026

Tabla A2.9: AL, CP %,L % and U % for several sample sizes and several resampling method of the estimators compared. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY $R(0,8,0,2)$

.

	AL	CP %	L %	U %	AL	CP %	L %	U %	AL	CP %	L %	U %	AL	CP %	L %	U %
									Booth et al., (1994)							
Estimator	$n = 100$				$n = 125$				$n = 150$				$n = 200$			
$\widehat{R}_{HT}$	1.5961	98.0392	1.9608	0	1.429	92.9204	6.1947	0.885	1.2533	95.1923	4.8077	0	1.0154	97.3913	2.6087	0
$\widehat{R}^1_{cal}$	1.5665	98.0392	0.9804	0.9804	1.378	93.8053	4.4248	1.7699	1.203	96.1538	0.9615	2.8846	0.9921	97.3913	0.8696	1.7391
$\widehat{R}^2_{cal}$	1.3911	98.0392	0.9804	0.9804	1.2221	96.4602	1.7699	1.7699	1.0724	97.1154	0.9615	1.9231	0.9042	97.3913	0.8696	1.7391
$\widehat{R}_{KL1cali}$	1.3863	99.0196	0	0.9804	1.2247	97.3451	0.885	1.7699	1.0784	94.2308	1.9231	3.8462	0.9052	95.6522	0.8696	3.4783
$\widehat{R}_{KL1calr}$	1.3906	99.0196	0	0.9804	1.2255	97.3451	0.885	1.7699	1.0778	94.2308	1.9231	3.8462	0.9054	95.6522	0.8696	3.4783
$\widehat{R}_{KL1calo}$	1.3908	99.0196	0	0.9804	1.2248	97.3451	0.885	1.7699	1.0777	94.2308	1.9231	3.8462	0.9055	95.6522	0.8696	3.4783
$\widehat{R}_{KL2cali}$	1.3865	99.0196	0	0.9804	1.2248	97.3451	0.885	1.7699	1.0784	94.2308	1.9231	3.8462	0.9051	95.6522	0.8696	3.4783
$\widehat{R}_{KL2calr}$	1.3853	99.0196	0	0.9804	1.2247	97.3451	0.885	1.7699	1.0781	94.2308	1.9231	3.8462	0.9052	95.6522	0.8696	3.4783
$\widehat{R}_{KL2calo}$	1.3853	99.0196	0	0.9804	1.2247	97.3451	0.885	1.7699	1.0781	94.2308	1.9231	3.8462	0.9052	95.6522	0.8696	3.4783
$\widehat{R}_{Dcali}$	1.4259	98.0392	0.9804	0.9804	1.2382	96.4602	1.7699	1.7699	1.0872	94.2308	1.9231	3.8462	0.9056	98.2609	0	1.7391
$\widehat{R}_{Dcalr}$	1.4252	98.0392	0.9804	0.9804	1.2382	96.4602	1.7699	1.7699	1.0871	94.2308	1.9231	3.8462	0.9055	98.2609	0	1.7391
$\widehat{R}_{Dcalo}$	1.4252	98.0392	0.9804	0.9804	1.2382	96.4602	1.7699	1.7699	1.0871	94.2308	1.9231	3.8462	0.9055	98.2609	0	1.7391
									Antal, E., Tillé, Y. (2014)							
$\widehat{R}_{HT}$	1.5582	98.0392	1.9608	0	1.4031	92.9204	5.3097	1.7699	1.2173	94.2308	5.7692	0	0.9974	96.5217	2.6087	0.8696
$\widehat{R}^1_{cal}$	1.6224	98.0392	0.9804	0.9804	1.4559	97.3451	0.885	1.7699	1.2358	95.1923	0.9615	3.8462	1.0085	96.5217	0	3.4783
$\widehat{R}^2_{cal}$	1.4064	98.0392	0.9804	0.9804	1.22	97.3451	0.885	1.7699	1.0981	97.1154	0	2.8846	0.8986	96.5217	0.8696	2.6087
$\widehat{R}_{KL1cali}$	1.3766	99.0196	0	0.9804	1.2206	97.3451	0.885	1.7699	1.0769	94.2308	0	5.7692	0.8907	96.5217	0.8696	2.6087
$\widehat{R}_{KL1calr}$	1.3743	99.0196	0	0.9804	1.2195	97.3451	0.885	1.7699	1.0776	94.2308	0	5.7692	0.8911	96.5217	0.8696	2.6087
$\widehat{R}_{KL1calo}$	1.3754	99.0196	0	0.9804	1.2196	97.3451	0.885	1.7699	1.0776	94.2308	0	5.7692	0.8908	96.5217	0.8696	2.6087
$\widehat{R}_{KL2cali}$	1.3767	99.0196	0	0.9804	1.2205	97.3451	0.885	1.7699	1.0771	94.2308	0	5.7692	0.8908	96.5217	0.8696	2.6087
$\widehat{R}_{KL2calr}$	1.3775	99.0196	0	0.9804	1.2204	97.3451	0.885	1.7699	1.0769	94.2308	0	5.7692	0.8909	96.5217	0.8696	2.6087
$\widehat{R}_{KL2calo}$	1.3774	99.0196	0	0.9804	1.2204	97.3451	0.885	1.7699	1.0771	94.2308	0	5.7692	0.891	96.5217	0.8696	2.6087
$\widehat{R}_{Dcali}$	1.4107	98.0392	0.9804	0.9804	1.2382	96.4602	1.7699	1.7699	1.0901	96.1538	0	3.8462	0.8964	95.6522	0.8696	3.4783
$\widehat{R}_{Dcalr}$	1.4117	98.0392	0.9804	0.9804	1.2381	96.4602	1.7699	1.7699	1.0903	96.1538	0	3.8462	0.8965	95.6522	0.8696	3.4783
$\widehat{R}_{Dcalo}$	1.4117	98.0392	0.9804	0.9804	1.2381	96.4602	1.7699	1.7699	1.0902	96.1538	0	3.8462	0.8965	95.6522	0.8696	3.4783
									Antal, E., Tillé, Y. (2011).							
$\widehat{R}_{HT}$	1.6156	96.0784	2.9412	0.9804	1.3913	93.8053	4.4248	1.7699	1.2315	96.1538	3.8462	0	1.0136	97.3913	2.6087	0
$\widehat{R}^1_{cal}$	1.5861	99.0196	0.9804	0	1.3981	92.0354	4.4248	3.5398	1.2276	97.1154	0.9615	1.9231	0.9898	97.3913	0	2.6087
$\widehat{R}^2_{cal}$	1.3979	98.0392	0.9804	0.9804	1.1907	97.3451	0.885	1.7699	1.0811	97.1154	0	2.8846	0.8968	94.7826	1.7391	3.4783
$\widehat{R}_{KL1cali}$	1.4045	99.0196	0	0.9804	1.2049	98.2301	0	1.7699	1.0862	95.1923	0.9615	3.8462	0.9001	94.7826	1.7391	3.4783
$\widehat{R}_{KL1calr}$	1.404	99.0196	0	0.9804	1.2027	98.2301	0	1.7699	1.0866	94.2308	0.9615	4.8077	0.8996	94.7826	1.7391	3.4783
$\widehat{R}_{KL1calo}$	1.4042	99.0196	0	0.9804	1.2031	98.2301	0	1.7699	1.0868	94.2308	0.9615	4.8077	0.8998	94.7826	1.7391	3.4783
$\widehat{R}_{KL2cali}$	1.4045	99.0196	0	0.9804	1.2051	98.2301	0	1.7699	1.0862	95.1923	0.9615	3.8462	0.9001	94.7826	1.7391	3.4783
$\widehat{R}_{KL2calr}$	1.4049	99.0196	0	0.9804	1.2049	98.2301	0	1.7699	1.0863	95.1923	0.9615	3.8462	0.9003	94.7826	1.7391	3.4783
$\widehat{R}_{KL2calo}$	1.4049	99.0196	0	0.9804	1.2049	98.2301	0	1.7699	1.0863	95.1923	0.9615	3.8462	0.9002	94.7826	1.7391	3.4783
$\widehat{R}_{Dcali}$	1.4496	98.0392	0.9804	0.9804	1.2182	96.4602	1.7699	1.7699	1.0949	97.1154	0	2.8846	0.9061	94.7826	1.7391	3.4783
$\widehat{R}_{Dcalr}$	1.4503	98.0392	0.9804	0.9804	1.2181	96.4602	1.7699	1.7699	1.0951	97.1154	0	2.8846	0.9063	94.7826	1.7391	3.4783
$\widehat{R}_{Dcalo}$	1.4503	98.0392	0.9804	0.9804	1.2181	96.4602	1.7699	1.7699	1.095	97.1154	0	2.8846	0.9063	94.7826	1.7391	3.4783

The results show a large decrease in bias and MSE for all ratio percentiles considered, for both calibration methods, and for the three versions of them based on linear, raking and logit response models, which shows the robustness of the adjustment method. Although the simulation results show that there is no uniformly better estimator than another among the proposed estimators (both with respect to bias and efficiency), the $\widehat{R}^{(2)}_{cal}$ and $\widehat{R}^{(3)}_{KL2cal}$ estimators are computationally simpler than the other alternatives which implies that they are a suitable option for the estimation of measures for wage inequality based on percentiles ratios.

Kott & Liao (2015) say that there are reasons for preferring the use of two calibration-weighting steps even when the sets of calibration variables used in both steps are the same or a subset of the calibration variables in a single step. These reasons, together with the good performance of the two-step estimator shown in the simulation study, suggest the choice of the estimator $\widehat{R}^{(2)}_{cal}$.

We used parametric methods to model the lack of response but we could use machine learning techniques as regression trees, spline regression, random forests etc. Other way to reduce the bias o is to combine calibration technique with other techniques as the Propensity Score Adjustment Ferri-García & Rueda (2018). Further research should 318 focus on extensions of those methods for general parameter estimation.

Acknowledgement

The work was supported by the Ministerio de Economia, Industria y Competitividad, Spain, under 315 Grant MTM2015-63609-R and by Ministerio de Ciencia e Innovación, Spain, under grant PID2019-106861RB-I00./ 316 10.13039/501100011033)

References

[1] Antal, E., & Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106(494), 534-543.

[2] Antal, E., & Tillé, Y. (2014). A new resampling method for sampling designs without replacement: the doubled half bootstrap. *Computational Statistics*, 29(5), 1345–1363.

[3] Booth, J. G., R. W. Butler, R. W., & Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89(428), 1282–1289.

[4] Burtless, G. (1999). Effects of growing wage disparities and changing family composition on the US income distribution. *European Economic Review*, 43(4-6), 853-865.

[5] Chambers, R.L., & Dunstan, A.(1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

[6] Chauvet, G. (2007). Bootstrap pour un tirage à plusieurs degrés avec échantillonnage à forte entropie à chaque degré, Working Papers 2007-39, Center for Research in Economics and Statistics.

[7] Deville, J. C. (2000). Generalized Calibration and Application to Weighting for Non-response. *COMPSTAT: Proceedings in Computational Statistics, 14th Symposium, Utrecht, The Netherlands*, eds. J. G. Bethlehem and P. G. M. van der Heijden, New York: Springer-Verlag, 65-76.

[8] Dickens, R., & Manning, A. (2004). Has the national minimum wage reduced UK wage inequality?. *Journal of the Royal Statistical Society: Series A*, 167(4), 613-626.

[9] Ferri-García, R., & Rueda, M. D. M. (2018). Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT*, 1(2), 159-162.

[10] Jones, A. F., & Weinberg, D. H. (2000). The changing shape of the nation's income distribution. *Current Population Reports*, 60, 1-11.

[11] Kott, P.S., & Liao, D. (2015). One step or two? Calibration weihting form a complete list frame with nonresponse. *Survey Methodolgy*, 41(1), 165-181.

[12] Kott, P.S., & Liao, D. (2017). Calibration weighting for nonresponse that is not missing at random: allowing more calibration than response-model variables. *Journal of Survey Statistics and Methodology*, 5,159–174.

[13] Machin, S., Manning, A., & Rahman, L. (2003). Where the minimum wage bites hard: introduction of minimum wages to a low wage sector. *Journal of the European Economic Association*, 1(1), 154-180.

[14] Rao, J.N.K., Kovar, J.G., & Mantel, H.J. (1990). On estimating distribution function and quantiles from survey data using auxiliary information. *Biometrika*, 77(2), 365-375.

[15] Rueda, M., Martínez-Puertas, S., & Illescas, M. (2021). Treating nonresponse in the estimation of the distribution function. *Mathematics and Computers in Simulation*, 186, 136-144.

[16] Rueda, M., Martínez, S., Martínez, H., & Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137(2), 435-448.

Apéndice A3

The optimization problem of quantile and poverty measures estimation based on calibration

Martínez, Sergio; Rueda, María del Mar; Illescas, María Dolores (2022)

The optimization problem of quantile and poverty measures estimation based on calibration.

Journal of Computational and Applied Mathematics, Vol. 405, pp. 113054.

DOI: 10.1016/j.cam.2020.113054

MATHEMATICS, APPLIED			
JCR Year	Impact factor	Rank	Quartile
2021	2.872	37/267	Q1

Abstract

New calibrated estimators of quantiles and poverty measures are proposed. These estimators combine the incorporation of auxiliary information provided by auxiliary variables related to the variable of interest by calibration techniques with the selection of optimal calibration points under simple random sampling without replacement. The problem of selecting calibration points that minimize the asymptotic variance of the quantile estimator is addressed. Once the problem is solved, the definition of the new quantile estimator requires that the optimal estimator of the distribution function on which it is based verifies the properties of the distribution function. Through a theorem, the nondecreasing monotony property for the optimal estimator of the distribution function is established and the corresponding optimal estimator can be defined. This optimal quantile estimator is also used to define new estimators for poverty measures. Simulation studies with real data from the Spanish living conditions survey compares the performance of the new estimators against various methods proposed previously, where some resampling techniques are used for the variance estimation. Based on the results of the simulation study, the proposed estimators show a good performance and are a reasonable alternative to other estimators.

1. Introduction

Quantile estimation is a issue of great interest because some measures and indicators depend on quantiles in many fields of research such as health science (Tellez-Plaza et al. (2008)); anthropology (Bogin & Sullivan (1986)) or economics (Decker et al. (2014)). More specifically, in the field of economics, studies on the analysis of poverty and social exclusion have an increasing importance for governments and society in general, since some poverty measures, like the proportion of people (or households) in poverty, are important measures of the country's overall economic welfare. Many indicators used in the poverty studies are based on quantiles, since they analyze variables with skewed distributions such as income, and in such cases the median is more suitable location measure than the mean. Thus, one of the commonly used measures in the poverty analysis is the poverty line that allows dividing the population into poor and nonpoor and that, for example, Eurostat fixes as 60 % of the median of the equivalent net income. Additionally, poverty studies incorporate the analysis of wage inequality and income distribution, whose measurement is often based on percentile ratios, such as 50th/5th and 50th/25th (Dickens & Manning (2004)); 50th/10th (Nickell (2004),Machin et al. (2003), Metcalf (2008), Burtless (1999)); 95th/50th (Machin et al. (2003), Burtless (1999)) and 90th/10th; 95th/20th; and 80th/20th (Jones & Weinberg (2000)).

In official surveys of living conditions, in social surveys and in sample surveys in general, auxiliary information is often available through additional variables related to the study variable. When auxiliary information is available, there are several alternative methods for incorporating it into the estimation phase and obtaining more efficient estimators (Deville & Särndal (1992); Chen & Sitter (1999); Chambers & Dunstan (1986)); Dorfman & Hall (1993)). These procedures have been applied to estimate the population mean (Rueda et al. (2006); Rueda et al. (2009)), the distribution function (Chambers & Dunstan (1986); Dorfman & Hall (1993); Singh et al. (2008); Mayor-Gallego et al. (2019)) quantiles (Harms & Duchesne (2006); Chen & Wu (2002)) and poverty measures Morales et al. (2018). Particularly, in the case of estimation of quantiles, the auxiliary information can be incorporated by means of indirect estimators. In this case, it is necessary to have the equivalent quantile of the auxiliary variable for a given quantile of the study variable (Kuk & Mak (1989); Rao et al. (1990)). Another possibility considers the incorporation of the auxiliary information to obtain estimators of the distribution function and to obtain the estimation of the quantile through the inverse function (Chambers & Dunstan (1986); Dorfman & Hall (1993)). This procedure requires that the estimator of the distribution function fulfills the distribution function's properties. Thus, based on this option, Rueda et al. (2007b) obtained quantile estimators based on calibration framework described in Rueda et al. (2007a). Similarly, also based on the same calibration framework, Martínez et al. (2011) developed post-stratified quantile estimators. The main advantage of the framework proposed in Rueda et al. (2007a) is that the obtained estimators are genuine distribution functions[1] under some conditions. One drawback of these estimators, is that their efficiency depends on the selection of some calibration points t_i. Recenlty, under simple random sampling, the problem of optimal selection points in order to obtain the best estimation is treated in Martínez et al. (2011)-Martínez et al. (2017). Unfortunately, the quantile estimation through the estimation of the distribution function needs the estimation for all value t and the optimal selection of auxiliary points depends on the point t in which we want to estimate the distribution function. This implies that the distribution function estimators based on optimal choice, in general, are not monotonous non-decreasing and may take values beyond the range $[0, 1]$.

In this work, we will adapt and employ the optimal selection proposals in Martínez et al. (2017) in the estimation of quantiles. We show that the problem of optimizing the variance of a quantile estimator is equivalent to the optimization of the variance of the distribution function estimator at one point. We demonstrate that under certain conditions, the estimators obtained through the optimal selection proposed in Martínez

[1]For an estimator $\hat{F}(t)$ of $F(t)$ to be a genuine distribution function it should be monotonic increasing and such that $\hat{F}(-\infty) = 0$ and $\hat{F}(+\infty) = 1$

et al. (2017) meet the distribution function properties and can be directly used in the quantile estimation. Due to the complexity of the quantile estimation and the optimal selection for calibration estimators, a practical mathematical expression for the variances of the quantile estimator could not be established. Thus, some resampling techniques will be employed to obtain variance estimation of the quantile estimators proposed. Finally, in this work we will define new percentile ratio estimators that can be applied in the estimation of poverty measures.

The remainder of the article is organized in four sections. After introducing the problem of quantile estimation in Section 2, in Section 3, new calibration quantiles estimators are proposed based on optimal selection points for the estimation of distribution function. In Section 4, we propose the use of resampling techniques for the variance estimation of the quantile estimators proposed in Section 4. The application of the optimal quantile estimators in poverty measures estimation is done in Section 5. Section 6 includes two simulation studies based on real survey data obtained from the Spanish living conditions survey in order to analyse the performance of quantile estimators and poverty measure estimators proposed in this work. Finally, Section 7 presents the concluding remarks.

2. Estimation of the distribution function and quantiles in survey sampling

Consider a finite population $U = \{1, \ldots, N\}$ with N different units where a sampling design $p(\cdot)$ is defined with first and second-order inclusion probabilities $\pi_k > 0$ and $\pi_{kl} > 0$ $k, l \in U$. A random sample $s = \{1, 2, \ldots, n\}$ of fixed size n is selected according to the sampling design $p(\cdot)$ and $d_k = \pi_k^{-1}$ denotes the sampling design-basic weight for unit $k \in U$ which is known. We denote by y_k the study variable and by x_k a vector of auxiliary variables at unit k. The values x_k are assumed to be known for all population units but the value y_k is assumed to be known only if the sample s includes the kth unit. The distribution function $F_y(t)$ of the study variable y is given by

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k) \tag{3.1}$$

where

$$\Delta(t - y_k) = \begin{cases} 1 & \text{si } t \geq y_k \\ 0 & \text{si } t < y_k. \end{cases}$$

Based on $F_y(t)$, the finite population α-quantile of y is defined as minimum value of t for which at least $100 \cdot \alpha\,\%$ of the y's values are less than or equal to that value, that is

$$Q_y(\alpha) = \text{ínf}\{t : F_y(t) \geq \alpha\} = F_y^{-1}(\alpha).$$

A general procedure to obtain an indirect estimator for $Q_y(\alpha)$ is based on the incorporation of auxiliary information in the estimation of $F_y(t)$ to obtain an estimator $\widehat{F}_y(t)$ that fulfills the distribution function's properties, that is, $\widehat{F}_y(t)$ is a genuine distribution function. Under this assumption, the quantile $Q_y(\alpha)$ can be estimated by taking the inverse of $\widehat{F}_y(t)$ in the following way:

$$\widehat{Q}_y(\alpha) = \text{ínf}\{t : \widehat{F}_y(t) \geq \alpha\} = \widehat{F}_y^{-1}(\alpha).$$

The usual estimator of the distribution function $F_y(t)$ is the Horvitz-Thompson estimator given by:

$$\widehat{F}_{YHT}(t) = \frac{1}{N}\sum_{k\in s} d_k \Delta(t - y_k). \tag{3.2}$$

The estimator $\widehat{F}_{YHT}(t)$ is unbiased and under simple random sampling, it verifies the distribution function properties, but generally it is not a genuine distribution function and does not use the auxiliary information provided by the vector x.

Recently, to incorporate the auxiliary information in the estimation of $F_y(t)$, some authors ((Harms & Duchesne (2006), Rueda et al. (2007a), Rueda et al. (2007b), Singh et al. (2008) and Arcos et al. (2017))) have used the calibration method in the estimation of the distribution function and quantiles. Specifically,Rueda et al. (2007a) modified the estimator $\widehat{F}_{YHT}(t)$ by the calibration method. To do so, they considered a pseudo-variable $g_k = \widehat{\beta}'\mathbf{x}_k$ for $k = 1, 2, ...N$, where

$$\widehat{\beta} = \left(\sum_{k\in s} d_k \mathbf{x}_k \mathbf{x}_k^{'}\right)^{-1} \cdot \sum_{k\in s} d_k \mathbf{x}_k y_k \tag{3.3}$$

and they replaced the basic weights d_k by new calibrated weights ω_k by means of the minimization of the chi-square distance measure

$$\Phi_s = \sum_{k\in s} \frac{(\omega_k - d_k)^2}{d_k q_k} \tag{3.4}$$

subject to the calibration equations

$$\frac{1}{N}\sum_{k\in s} \omega_k \Delta(t_j - g_k) = F_g(t_j) \quad j = 1, 2, \ldots, P \tag{3.5}$$

where q_k are known positive constants unrelated to d_k, $F_g(t_j)$ denotes the finite distribution function of the pseudo-variable g_k evaluated at the points $t_j \quad j = 1, 2, \ldots, P$ and it is assumed, with no loss in generality, that $t_1 < t_2 < \ldots t_P$.

The resulting estimator (Rueda et al. (2007a)) is given by

$$\widehat{F}_{yc}(t) = \widehat{F}_{YHT}(t) + \left(F_g(\mathbf{t_g}) - \widehat{F}_{GHT}(\mathbf{t_g})\right)' \cdot \widehat{D}(\mathbf{t_g}) \tag{3.6}$$

where $\widehat{F}_{GHT}(\mathbf{t_g})$ is the Horvitz-Thompson estimator of $F_g(\mathbf{t_g})$ evaluated at $\mathbf{t_g} = (t_1, \ldots, t_P)'$ and

$$\widehat{D}(\mathbf{t_g}) = T^{-1} \cdot \sum_{k \in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(t - y_k)$$

assuming that the matrix T, given by

$$\sum_{k \in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)'$$

is nonsingular.

Under some conditions (Rueda et al. (2007a)) the estimator $\widehat{F}_{yc}(t)$ is a genuine distribution function and based on this framework, Rueda et al. (2007b) developed a new estimator for quantiles $Q_y(\alpha)$.

3. Optimal quantile estimators based on calibration estimation

In this section we will consider the search for quantile calibration estimators that are optimal in the sense of least error.

3.1. The optimization problem

A quantile estimator $\widehat{Q}_y(\alpha)$ can be expressed asymptotically as a linear function of the estimated distribution function evaluated at the quantile $Q_y(\alpha)$ by the Bahadur representation (see Chambers & Dunstan (1986)):

$$\widehat{Q}_y = \frac{1}{f_y(Q_y(\alpha))}(\alpha - \widehat{F}_y(Q_y(\alpha))) + O(n^{-1/2}), \tag{3.7}$$

where $f_y(\cdot)$ denotes the derivative of the limiting value of $F_y(\cdot)$ as $N \longrightarrow \infty$. This linear approximation previously used by Kuk & Mak (1989) and Arcos et al. (2007) helps to study the asymptotic properties of the estimator. Using this approximation we can express the asymptotic variance of $\widehat{Q}_y(\alpha)$ as

$$V_{asym}(\widehat{Q}_y(\alpha)) = \left(\frac{1}{f_y(Q_y(\alpha))}\right)^2 V(\widehat{F}_y(Q_y(\alpha)))$$

then the problem of minimizing the variance of the quantile's estimator $Q_y(\alpha)$ is the same as minimizing the variance of the estimator of the distribution function $\widehat{F}_y(Q_y(\alpha))$ on which it is based. Since the value $Q_y(\alpha)$ is unknown, it is not possible to obtain the optimal points for the estimation of $\widehat{F}_{yc}(Q_y(\alpha))$ following the approach developed in Martínez et al. (2017). Consequently, for the optimal estimate of $Q_y(\alpha)$, we consider the optimal estimation of $F_y(t)$ for each point t.

Following Rueda et al. (2007a), the asymptotic variance of $\widehat{F}_{yc}(t)$ is given by:

$$AV(\widehat{F}_{yc}(t)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl}(d_k E_k)(d_l E_l) \tag{3.8}$$

where $E_k = \Delta(t - y_k) - \Delta(\mathbf{t_g} - g_k) \cdot D(\mathbf{t_g})$, with

$$D(\mathbf{t_g}) = \left(\sum_{k \in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)' \right)^{-1} \cdot \left(\sum_{k \in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(t - y_k) \right).$$

Thus, the selection of the auxiliary vector $\mathbf{t_g}$ changes the precision of the calibration estimator $\widehat{Q}_y(\alpha)$.

Following Martínez et al. (2015), under simple random sampling without replacement, the minimization of asymptotic variance (4.8) is equivalent to the minimization of the function:

$$Q_t(\gamma_1, \ldots, \gamma_P) = 2NF_y(t) \cdot K_t(\gamma_P) - \sum_{j=1}^{P} \frac{\left(K_t(\gamma_j) - K_t(\gamma_{j-1})\right)^2}{(F_g(\gamma_j) - F_g(\gamma_{j-1}))} - \left(K_t(\gamma_P)\right)^2$$

with $K_t(\gamma) = \sum_{k \in U} \Delta(\gamma - g_k)\Delta(t - y_k)$.

Under simple random sampling without replacement, the function $Q_t(\gamma_1, \ldots, \gamma_P)$ has its minimum at a vector $\mathbf{t_P} = (\gamma_{t1}, \ldots, \gamma_{tP})$ with $\gamma_{tj} \in A_t \cup B_t$, $j = 1, \ldots, P$ where

$$A_t = \{g_k : k \in U; y_k \leq t\} = \{a_1^t, a_2^t, \ldots, a_{M_t}^t\} \quad \text{with} \quad a_h^t < a_{h+1}^t \quad \text{for} \quad h = 1, \ldots, M_t - 1 \tag{3.9}$$

where M_t is the number of elements in the set A_t and

$$B_t = \{b_1^t, b_2^t, \ldots, b_M^t\}$$

with

$$b_1^t = \text{máx}_{l \in U_1}\{g_l\} \quad \text{where } U_1 = \{l \in U : g_l < a_1^t\}$$
$$b_h^t = \text{máx}_{l \in U_h}\{g_l\} \quad \text{where } U_h = \{l \in U : a_{h-1}^t \leq g_l < a_h^t\} \quad h = 2, 3, \ldots, M_t$$

and $b_h^t \leq b_{h+1}^t$ for $h = 1, \ldots, M_t - 1$.

Under simple random sampling without replacement Martínez et al. (2017) found that the auxiliary vector $\mathbf{t}_g$ has optimal dimension $P = 2m_t$ when b_1^t exists and for all $j = 2, \ldots, m_t$, $b_j^t \neq a_{j-1}^t$ and the optimal vector is given by

$$\mathbf{t_{OPT}}(t) = (b_1^t, a_1^t, \ldots, b_{m_t}^t, a_{M_t}^t). \tag{3.10}$$

In the case that for some values $j_1^t, j_2^t, \ldots j_{p_t}^t \in \{1, \ldots, m_t\}$; $a_{j_1-1}^t = b_{j_1}^t$ with $p_t \leq m_t$ and $j_h^t \neq j_q^t$ if $h \neq q$ the optimal dimension is given by $P = 2m_t - p_t$ and the optimal auxiliary vector $\mathbf{t_{OP}}$ is:

$$\mathbf{t_{OP}}(t) = (b_1^t, a_1^t, b_2^t, a_2^t, \ldots, b_{j_1-1}^t, a_{j_1-1}^t, a_{j_1}^t, b_{j_1+1}^t, \ldots, b_{j_h-1}^t, a_{j_h-1}^t, a_{j_h}^t, b_{j_h+1}^t, \ldots b_{M_t}^t, a_{M_t}^t). \tag{3.11}$$

Generally, the optimal vector $\mathbf{t_{OPT}}(t)$ is unknown. Moreover, if its value is known, it can produce some problems when it is used with the data of a particular sample s (it can produce incompatible calibration restrictions in (4.6)). Thus, in a similar way to the previous cases, we consider a estimated vector $\widehat{\mathbf{t}}_{\mathbf{OP}}(t)$ based on the set A_{st} and B_{st} defined as:

$$A_{st} = \{g_k : k \in s; y_k \leq t\} = \{a_1^t, a_2^t, \ldots, a_{m_t}^t\}$$

with $a_h^t < a_{h+1}^t$ for $h = 1, \ldots, m_t - 1$ and B_{st} is defined, based on the sample s, in a similar way that B_t.

Then we define the calibration estimator for the distribution function estimator:

$$\widehat{F}_{YO}(t) = \widehat{F}_{YHT}(t) + \left(F_g(\widehat{\mathbf{t}}_{\mathbf{OP}}) - \widehat{F}_{GHT}(\widehat{\mathbf{t}}_{\mathbf{OP}})\right)' \cdot \widehat{D}(\widehat{\mathbf{t}}_{\mathbf{OP}}) \tag{3.12}$$

where

$$\widehat{D}(\widehat{\mathbf{t}}_{\mathbf{OP}}) = \frac{\sum_{k \in s} d_k q_k \Delta(\widehat{\mathbf{t}}_{\mathbf{OP}} - g_k)\Delta(t - y_k)}{\sum_{k \in s} d_k q_k \Delta(\widehat{\mathbf{t}}_{\mathbf{OP}} - g_k)}.$$

Since the optimal vector $\widehat{\mathbf{t}}_{\mathbf{OP}}$ depends on t, the estimator $\widehat{F}_{YO}(t)$ considers different calibration equations for each value of t. Consequently, the conditions developed in Rueda et al. (2007a) for $\widehat{F}_{yc}(t)$ in general do not guarantee that $\widehat{F}_{YO}(t)$ is a genuine distribution function. In the next subsection we will see that $\widehat{F}_{YO}(t)$

meets the conditions of a true distribution function.

3.2. Defining the optimal quantile estimator.

In order to define the optimal quantile estimator, we must first demonstrate that the estimator $\widehat{F}_{YO}(t)$ is a genuine distribution function and a key property is nondecreasing monotony property. We consider the usual weights $q_k = 1$ (the uniform weighting is likely to dominate in applications Deville & Särndal (1992)). The following theorem establish the nondecreasing monotony property for $\widehat{F}_{YO}(t)$.

Theorem. The calibration estimator $\widehat{F}_{YO}(t)$ is monotone nondecreasing.

Proof:

If we consider values $t \leq z$ with $y_{[i]} \leq t \leq z < y_{[i+1]}$, and we denote by $B_{si} = B_{sy_{[i]}}$, it is clear that

$$A_{st} = A_{sz} = A_{si} \quad \text{and} \quad B_{st} = B_{sz} = B_{si}.$$

Consequently, $\widehat{\mathbf{t}}_{\mathbf{OP}}(t) = \widehat{\mathbf{t}}_{\mathbf{OP}}(z) = \widehat{\mathbf{t}}_{\mathbf{OP}}(y[i])$ and calibration weights ω_k in (4.6) are obtained with the same auxiliary vector for t and z and following (Rueda et al. (2007a)), since $q_k = 1$ for all $k \in s$, we have $\widehat{F}_{YO}(t) \leq \widehat{F}_{YO}(z)$.

Now, we consider the case where $t \leq z$ with $y_{[i]} \leq t < y_{[i+1]}$ and $y_{[i+1]} \leq z < y_{[i+2]}$; $i = 1, \ldots, l-2$. For $y_{[i]} \leq t < y[i+1]$, we have:

$$A_{si} = \{a^i_1, \ldots, a^i_{m_i}\} \quad ; \quad B_{si} = \{b^i_1, \ldots, b^i_{m_i}\}.$$

We denote by $R_{si} = \{j : b^i_j = a^i_{j-1}\}$ and $\bar{R}_{si} = \{j : b^i_j \neq a^i_{j-1}\}$. It is clear that $\{1, \ldots, m_i\} = R_{si} \cup \bar{R}_{si}$. Now, if we assume that $R_{si} = \emptyset$, then the optimal vector $\widehat{\mathbf{t}}_{\mathbf{OPT}}(t)$ is given by the sample-based version of (4.12) and following (Rueda et al. (2007a)), the calibration estimator $\widehat{F}_{YO}(t)$ is given by:

$$\widehat{F}_{YO}(t) = \widehat{F}_{YHT}(t) + \sum_{j=1}^{2m_i} \left(F_g(t_j) - \widehat{F}_{GHT}(t_j) \right) \cdot A_i(t_j) \tag{3.13}$$

with

$$A_i(a^i_j) = \frac{\sum_{k \in s} d_k \Delta(a^i_j - g_k)\Delta(t - y_k) - \sum_{k \in s} d_k \Delta(b^i_j - g_k)\Delta(t - y_k)}{N\left(\widehat{F}_{GHT}(a^i_j) - \widehat{F}_{GHT}(b^i_j)\right)}$$

$$- \frac{\sum_{k \in s} d_k \Delta(b^i_{j+1} - g_k)\Delta(t - y_k) - \sum_{k \in s} d_k \Delta(a^i_j - g_k)\Delta(t - y_k)}{N\left(\widehat{F}_{GHT}(b^i_{j+1}) - \widehat{F}_{GHT}(a^i_j)\right)} =$$

$$\frac{\left(\widehat{k}_i(a^i_j) - \widehat{k}_i(b^i_j)\right)}{N\left(\widehat{F}_{GHT}(a^i_j) - \widehat{F}_{GHT}(b^i_j)\right)} = \frac{\left(\widehat{k}_i(a^i_j) - \widehat{k}_i(a^i_{j-1})\right)}{N\left(\widehat{F}_{GHT}(a^i_j) - \widehat{F}_{GHT}(b^i_j)\right)}; \quad j = 1, \ldots, m_i \tag{3.14}$$

$$A_i(b^i_j) = \frac{\sum_{k\in s} d_k \Delta(b^i_j - g_k)\Delta(t - y_k) - \sum_{k\in s} d_k \Delta(a^i_{j-1} - g_k)\Delta(t - y_k)}{N\left(\widehat{F}_{GHT}(b^i_j) - \widehat{F}_{GHT}(a^i_{j-1})\right)}$$

$$-\frac{\sum_{k\in s} d_k \Delta(a^i_j - g_k)\Delta(t - y_k) - \sum_{k\in s} d_k \Delta(b^i_j - g_k)\Delta(t - y_k)}{N\left(\widehat{F}_{GHT}(a^i_j) - \widehat{F}_{GHT}(b^i_j)\right)} = -A_i(a^i_j); \quad j = 1, \ldots, m_i \tag{3.15}$$

since it's easy to see that $\widehat{k}_i(a^i_{j-1}) = \widehat{k}_i(b^i_j)$ ($\widehat{k}$ is defined similarly to K but based on sample s) and where $\widehat{k}_i(a^i_0) = 0$ and $\widehat{k}_i(b^i_{m_i+1}) = \widehat{k}_i(a^i_{m_i})$ as we consider $a^i_0 < \min\{g_k : k \in U\}$ and $b_{m_i+1} > g_M$.
By replacing the values $A_i(a^i_j)$ and $A_i(b^i_j)$ in the equation (3.13), it could be easily seen how the estimator $\widehat{F}_{YO}(t)$ for $y_{[i]} \le t < y_{[i+1]}$ takes the following expression:

$$\widehat{F}_{YO}(t) = \sum_{j=1}^{m_i} \frac{\left(F_g(a^i_j) - F_g(b^i_j)\right) \cdot \left(\widehat{k}_i(a^i_j) - \widehat{k}_i(a^i_{j-1})\right)}{N \cdot \left(\widehat{F}_{GHT}(a^i_j) - \widehat{F}_{GHT}(b^i_j)\right)}. \tag{3.16}$$

Now, if we suppose that $\bar{R}_{si} = \emptyset$, then the optimal vector $\mathbf{t_{OPT}}(t) = (a^i_1, a^i_2, \ldots, a^i_{m_i})$ and $\widehat{F}_{YO}(t)$ take the following expression:

$$\widehat{F}_{YO}(t) = \widehat{F}_{YHT}(t) + \sum_{j=1}^{m_i} \left(F_g(a^i_j) - \widehat{F}_{GHT}(a^i_{j-1})\right) \cdot A_i(a^i_j) \tag{3.17}$$

where for $j = 1, \ldots, m_i - 1$

$$A_i(a^i_j) = \frac{\left(\widehat{k}_i(a^i_j) - \widehat{k}_i(a^i_{j-1})\right)}{N\left(\widehat{F}_{GHT}(a^i_j) - \widehat{F}_{GHT}(a^i_{j-1})\right)} - \frac{\left(\widehat{k}_i(a^i_{j+1}) - \widehat{k}_i(a^i_j)\right)}{N\left(\widehat{F}_{GHT}(a^i_{j+1}) - \widehat{F}_{GHT}(a^i_j)\right)}. \tag{3.18}$$

From (3.17) and (3.18), the calibration estimator $\widehat{F}_{YO}(t)$ for $y_{[i]} \le t < y_{[i+1]}$ is:

$$\widehat{F}_{YO}(t) = \sum_{j=1}^{m_i} \frac{\left(F_g(a^i_j) - F_g(a^i_{j-1})\right) \cdot \left(\widehat{k}_i(a^i_j) - \widehat{k}_i(a^i_{j-1})\right)}{N \cdot \left(\widehat{F}_{GHT}(a^i_j) - \widehat{F}_{GHT}(a^i_{j-1})\right)}. \tag{3.19}$$

Finally, we consider the case where $R_{si} \neq \emptyset$ and $\bar{R}_{si} \neq \emptyset$.
Be $R_{si} = \{j_1, \ldots j_{p_i}\}$, for all $j_h \in R_{si}$ we have:

$$A_i(a^i_{j_h}) = \frac{\left(\widehat{k}_i(a^i_{j_h}) - \widehat{k}_i(a^i_{j_h-1})\right)}{N\left(\widehat{F}_{GHT}(a^i_{j_h}) - \widehat{F}_{GHT}(a^i_{j_h-1})\right)}. \tag{3.20}$$

For $j \in \bar{R}_{si}$ and $j \neq j_h - 1$ for all $j_h \in R_{si}$, $A_i(a^i_j)$ and $A_i(b^i_j)$ are given by (3.14) and (3.15) while for $j = j_h - 1$ with $j \in \bar{R}_{si}$, $A_i(a^i_j)$ and $A_i(b^i_j)$ are given by (3.18) and (3.15).

Thus, in this case, the estimator $\widehat{F}_{YO}(t)$ for $y_{[i]} \leq t < y_{[i+1]}$ is given by:

$$\widehat{F}_{YO}(t) = \sum_{j \in R_{si}} \frac{\left(F_g(a^i_j) - F_g(a^i_{j-1})\right) \cdot \left(\widehat{k}_i(a^i_j) - \widehat{k}_i(a^i_{j-1})\right)}{N \cdot \left(\widehat{F}_{GHT}(a^i_j) - \widehat{F}_{GHT}(a^i_{j-1})\right)} +$$

$$\sum_{j \in \bar{R}_{si}} \frac{\left(F_g(a^i_j) - F_g(b^i_j)\right) \cdot \left(\widehat{k}_i(a^i_j) - \widehat{k}_i(a^i_{j-1})\right)}{N \cdot \left(\widehat{F}_{GHT}(a^i_j) - \widehat{F}_{GHT}(b^i_j)\right)}. \tag{3.21}$$

Now, for $y_{[i+1]} \leq z < y_{[i+2]}$, we consider the sets $A_{s(i+1)}$ and $B_{s(i+1)}$.

$$A_{s(i+1)} = \{a^{i+1}_1, \ldots, a^{i+1}_{m_{i+1}}\} \quad ; \quad B_{s(i+1)} = \{b^{i+1}_1, \ldots, b^{i+1}_{m_{i+1}}\}$$

and we define similarly the sets $R_{s(i+1)}$ and $\bar{R}_{s(i+1)}$.

Let be $A_{si} = A_{s(i+1)}$, then we have $B_{si} = B_{s(i+1)}$ and $\widehat{\mathbf{t}}_{\mathbf{OP}}(t) = \widehat{\mathbf{t}}_{\mathbf{OP}}(z) = \widehat{\mathbf{t}}_{\mathbf{OP}}(y[i])$. As in the previous case, we have $\widehat{F}_{YO}(t) \leq \widehat{F}_{YO}(z)$ because for both values t and z the weights ω_k in (4.6) are obtained with the same auxiliary vector.

If we assume that $A_{si} \neq A_{s(i+1)}$, because $A_{si} \subset A_{s(i+1)}$ then exist a set

$$H_{si} = \{r_h : h = 1, \ldots, m_i\} \subset \{j : j = 1, \ldots m_{i+1}\}$$

such that

$$a^i_1 = a^{i+1}_{r_1}; \ldots; a^i_{m_i} = a^{i+1}_{r_{m_i}} \tag{3.22}$$

with $r_1 < r_2 < \cdots < r_{m_i}$ and $a^{i+1}_{r_{(h-1)}} \leq a^{i+1}_{r_h - 1}$ for all $h \in \{1, \ldots, m_i\}$.

We denote by $\bar{H}_{si}$ the following set

$$\bar{H}_{si} = \{j : j = 1, \ldots, m_{i+1}\} - H_{si}.$$

On the other hand, since $a^{i+1}_{r_{(h-1)}} \leq a^{i+1}_{r_h - 1}$ for all $h = 1, \ldots, m_t$

$$\{g_k : a^i_{h-1} \leq g_k < a^i_h\} = \{g_k : a^{i+1}_{r_{(h-1)}} \leq g_k < a^{i+1}_{r_h}\} =$$

$$\{g_k : a^{i+1}_{r_{(h-1)}} \leq g_k \leq a^{i+1}_{r_h - 1}\} \cup \{g_k : a^{i+1}_{r_h - 1} \leq g_k < a^{i+1}_{r_h}\}$$

and therefore for all $h \in \{1, \ldots, m_i\}$

$$b_h^i = \max\{g_k : a_{h-1}^i \leq g_k < a_h^i\} = \max\{g_k : a_{r_h-1}^{i+1} \leq g_k < a_{r_h}^{i+1}\} = b_{r_h}^{i+1}. \tag{3.23}$$

By (3.22) and (3.23), then we have

$$\widehat{k}_i(a_{r_h}^{i+1}) = \widehat{k}_i(a_h^i) \quad ; \quad \widehat{k}_i(a_{r_{(h-1)}}^{i+1}) = \widehat{k}_i(a_{h-1}^i) = \widehat{k}_i(b_h^i) = \widehat{k}_i(b_{r_h}^{i+1}). \tag{3.24}$$

Now, we define Γ_h as:

$$\Gamma_h = \left(\widehat{k}_{i+1}(a_{r_h}^{i+1}) - \widehat{k}_{i+1}(a_{r_h-1}^{i+1})\right) - \left(\widehat{k}_i(a_{r_h}^{i+1}) - \widehat{k}_i(a_{r_{(h-1)}}^{i+1})\right) \tag{3.25}$$

where

$$\widehat{k}_{i+1}(a_{r_h}^{i+1}) = \sum_{k \in s} d_k \Delta(a_{r_h}^{i+1} - g_k)\Delta(y_{[i+1]} - y_k) =$$

$$\widehat{k}_i(a_{r_h}^{i+1}) + \sum_{k \in s} d_k \Delta(a_{r_h}^{i+1} - g_k) I_{[i+1]}(y_k)$$

with

$$I_{[i]}(y_k) = \begin{cases} 0 & \text{if } y_k \neq y_{[i]} \\ 1 & \text{if } y_k = y_{[i]}. \end{cases}$$

We denote by $\widehat{q}_{i+1}(z) = \sum_{k \in s} d_k \Delta(z - g_k) I_{[i+1]}(y_k)$. Thus $\widehat{k}_{i+1}(a_{r_h}^{i+1}) = \widehat{k}_i(a_{r_h}^{i+1}) + \widehat{q}_{i+1}(a_{r_h}^{i+1})$.

Similarly

$$\widehat{k}_{i+1}(a_{r_h-1}^{i+1}) = \widehat{k}_{i+1}(b_{r_h}^{i+1}) = \widehat{k}_i(b_{r_h}^{i+1}) + \widehat{q}_{i+1}(b_{r_h}^{i+1}) = \widehat{k}_i(a_{r_{(h-1)}}^{i+1}) + \widehat{q}_{i+1}(b_{r_h}^{i+1}).$$

Since $b_{r_h}^{i+1} < a_{r_h}^{i+1}$ for all $h \in \{1, \ldots, m_i\}$

$$\Gamma_h = \left(\widehat{k}_{i+1}(a_{r_h}^{i+1}) - \widehat{k}_{i+1}(a_{r_h-1}^{i+1})\right) - \left(\widehat{k}_i(a_{r_h}^{i+1}) - \widehat{k}_i(a_{r_{(h-1)}}^{i+1})\right) = \widehat{q}_{i+1}(a_{r_h}^{i+1}) - \widehat{q}_{i+1}(b_{r_h}^{i+1}) \geq 0. \tag{3.26}$$

Based on the sets R_{si}, $\bar{R}_{si}$, $R_{s(i+1)}$ and $\bar{R}_{s(i+1)}$ we consider several cases:

Case 1) $R_{si} = \emptyset$.

In this case, if we assume and $R_{s(i+1)} \neq \emptyset$ and $\bar{R}_{s(i+1)} \neq \emptyset$, the set $\{j : j = 1, \ldots, m_{i+1}\}$ is given by:

$$\{j : j = 1, \ldots, m_{i+1}\} = C_1 \cup C_2 \cup C_3 \cup C_4 =$$

$$= (R_{s(i+1)} \cap H_{si}) \cup (R_{s(i+1)} \cap \bar{H}_{si}) \cup (\bar{R}_{s(i+1)} \cap H_{si}) \cup (\bar{R}_{s(i+1)} \cap \bar{H}_{si}).$$

For $h = 1, \ldots, m_i$ with $r_h \in C_1 = R_{s(i+1)} \cap H_{si}$

$$a^{i+1}_{r_h-1} = b^{i+1}_{r_h} = b^i_h \neq a^i_{h-1} = a^{i+1}_{r_{(h-1)}} \tag{3.27}$$

while for $h = 1, \ldots, m_i$ with $r_h \in C_3 = \bar{R}_{s(i+1)} \cap H_{si}$

$$a^{i+1}_{r_h-1} \neq b^{i+1}_{r_h} = b^i_h \neq a^i_{h-1} = a^{i+1}_{r_{(h-1)}}. \tag{3.28}$$

By (3.16); (3.27) and (3.28), $\widehat{F}_{YO}(t)$ is given by:

$$\widehat{F}_{YO}(t) = \sum_{h=1}^{m_i} \frac{\left(F_g(a^i_h) - F_g(b^i_h)\right)}{\left(\widehat{F}_{GHT}(a^i_h) - \widehat{F}_{GHT}(b^i_h)\right)} \cdot \frac{\left(\widehat{k}_i(a^i_h) - \widehat{k}_i(a^i_{h-1})\right)}{N} =$$

$$= \sum_{\substack{h=1 \\ r_h \in C_1}}^{m_i} \frac{\left(F_g(a^{i+1}_{r_h}) - F_g(a^{i+1}_{r_h-1})\right) \cdot \left(\widehat{k}_i(a^{i+1}_{r_h}) - \widehat{k}_i(a^{i+1}_{r_{(h-1)}})\right)}{N \cdot \left(\widehat{F}_{GHT}(a^{i+1}_{r_h}) - \widehat{F}_{GHT}(a^{i+1}_{r_h-1})\right)} +$$

$$\sum_{\substack{h=1 \\ r_h \in C_3}}^{m_i} \frac{\left(F_g(a^{i+1}_{r_h}) - F_g(b^{i+1}_{r_h})\right) \cdot \left(\widehat{k}_i(a^{i+1}_{r_h}) - \widehat{k}_i(a^{i+1}_{r_{(h-1)}})\right)}{N \cdot \left(\widehat{F}_{GHT}(a^{i+1}_{r_h}) - \widehat{F}_{GHT}(b^{i+1}_{r_h})\right)}. \tag{3.29}$$

From (3.21), $\widehat{F}_{YO}(z)$ takes the following expression:

$$\widehat{F}_{YO}(z) = \sum_{r_h \in C_1} \frac{\left(F_g(a^{i+1}_{r_h}) - F_g(a^{i+1}_{r_h-1})\right) \cdot \left(\widehat{k}_{i+1}(a^{i+1}_{r_h}) - \widehat{k}_{i+1}(a^{i+1}_{r_h-1})\right)}{N \cdot \left(\widehat{F}_{GHT}(a^{i+1}_{r_h}) - \widehat{F}_{GHT}(a^{i+1}_{r_h-1})\right)} +$$

$$+ \sum_{r_h \in C_3} \frac{\left(F_g(a^{i+1}_{r_h}) - F_g(b^{i+1}_{r_h})\right) \cdot \left(\widehat{k}_{i+1}(a^{i+1}_{r_h}) - \widehat{k}_{i+1}(a^{i+1}_{r_h-1})\right)}{N \cdot \left(\widehat{F}_{GHT}(a^{i+1}_{r_h}) - \widehat{F}_{GHT}(b^{i+1}_{r_h})\right)} + W_s(z) \tag{3.30}$$

where $W_s(z) = V_s(z) + T_s(z)$ and

$$V_s(z) = \sum_{j \in C_2} \frac{\left(F_g(a^{i+1}_j) - F_g(a^{i+1}_{j-1})\right) \cdot \left(\widehat{k}_{i+1}(a^{i+1}_j) - \widehat{k}_{i+1}(a^{i+1}_{j-1})\right)}{N \cdot \left(\widehat{F}_{GHT}(a^{i+1}_j) - \widehat{F}_{GHT}(a^{i+1}_{j-1})\right)} \geq 0$$

$$T_s(z) = \sum_{j \in C_4} \frac{\left(F_g(a_j^{i+1}) - F_g(b_j^{i+1})\right) \cdot \left(\widehat{k}_{i+1}(a_j^{i+1}) - \widehat{k}_{i+1}(a_{j-1}^{i+1})\right)}{N \cdot \left(\widehat{F}_{GHT}(a_j^{i+1}) - \widehat{F}_{GHT}(b_j^{i+1})\right)} \geq 0.$$

Consequently, from (3.30) and (3.29), we have

$$\widehat{F}_{YO}(z) - \widehat{F}_{YO}(t) = W_s(z) + \sum_{r_h \in C_1} \frac{\left(F_g(a_{r_h}^{i+1}) - F_g(a_{r_h-1}^{i+1})\right) \cdot \Gamma_h}{N \cdot \left(\widehat{F}_{GHT}(a_{r_h}^{i+1}) - \widehat{F}_{GHT}(a_{r_h-1}^{i+1})\right)} +$$

$$\sum_{r_h \in C_3} \frac{\left(F_g(a_{r_h}^{i+1}) - F_g(b_{r_h}^{i+1})\right) \cdot \Gamma_h}{N \cdot \left(\widehat{F}_{GHT}(a_{r_h}^{i+1}) - \widehat{F}_{GHT}(b_{r_h}^{i+1})\right)} \geq 0.$$

On the other hand, if we suppose that $R_{s(i+1)} = \emptyset$, then $C_1 = C_2 = \emptyset$. For all $h \in \{1, \ldots, m_i\}$, we have (3.28) and $\widehat{F}_{YO}(t)$ is given by (3.29) with the sum in C_1 null. From (3.16), $\widehat{F}_{YO}(z)$ is given by (3.30) with $V_s(z) = 0$ and the summation based on the set C_1 null. Similarly, we can see that $\widehat{F}_{YO}(z) - \widehat{F}_{YO}(t) \geq 0$.
Identically, it can be shown that $\widehat{F}_{YO}(z) - \widehat{F}_{YO}(t) \geq 0$ in the case that $\bar{R}_{s(i+1)} = \emptyset$. From (3.19) (3.27) and because the sets C_3 and C_4 are empty, $\widehat{F}_{YO}(t)$ and $\widehat{F}_{YO}(z)$ are given by (3.29) and(3.30) respectively, with the summations based on null C_3 and $T_s(z) = 0$.

Case 2) $R_{si} \neq \emptyset$ and $\bar{R}_{si} \neq \emptyset$

For $h = 1, \ldots, m_i$ with $h \in R_{si}$,

$$a_h^i = a_{r_h}^{i+1} \quad ; \quad b_{r_h}^{i+1} = b_h^i = a_{h-1}^i = a_{r_{(h-1)}}^{i+1}$$

since $a_{r_{(h-1)}}^{i+1} \leq a_{r_h-1}^{i+1} \leq b_{r_h}^{i+1} = a_{r_{(h-1)}}^{i+1}$, we have

$$a_{r_h-1}^{i+1} = b_{r_h}^{i+1} = b_h^i = a_{h-1}^i = a_{r_{(h-1)}}^{i+1} \tag{3.31}$$

and consequently, for $h = 1, \ldots, m_i$ with $h \in R_{si}$; $r_h \in R_{s(i+1)} \cap H_{si} = C_1$ and the set $R_{s(i+1)} \neq \emptyset$. Because in this case $R_{s(i+1)} \neq \emptyset$ if we assume that $\bar{R}_{s(i+1)} \neq \emptyset$, then as in Case 1) with $R_{s(i+1)} \neq \emptyset$ and $\bar{R}_{s(i+1)} \neq \emptyset$, the value $\widehat{F}_{YO}(z)$ is given by (3.30).
Additionally, for $h = 1, \ldots, m_i$ with $h \in \bar{R}_{si}$, it is clear that $r_h \in C_1 \cup C_3$. For $h \in \bar{R}_{si}$ with $r_h \in C_1$ condition (3.27) is verified while for $h \in \bar{R}_{si}$ with $r_h \in C_3$; condition (3.28) is satisfied.

From (3.21); (3.31); (3.27) and (3.28); the value $\widehat{F}_{YO}(t)$ takes the following expression:

$$\widehat{F}_{YO}(t) = \sum_{h\in R_{si}} \frac{\left(F_g(a^{i+1}_{r_h}) - F_g(a^{i+1}_{r_h-1})\right)\cdot\left(\widehat{k}_i(a^{i+1}_{r_h}) - \widehat{k}_i(a^{i+1}_{r_{(h-1)}})\right)}{N\cdot\left(\widehat{F}_{GHT}(a^{i+1}_{r_h}) - \widehat{F}_{GHT}(a^{i+1}_{r_h-1})\right)} +$$

$$+ \sum_{\substack{h\in \bar{R}_{si}\\ r_h\in C_1}} \frac{\left(F_g(a^{i+1}_{r_h}) - F_g(a^{i+1}_{r_h-1})\right)\cdot\left(\widehat{k}_i(a^{i+1}_{r_h}) - \widehat{k}_i(a^{i+1}_{r_{(h-1)}})\right)}{N\cdot\left(\widehat{F}_{GHT}(a^{i+1}_{r_h}) - \widehat{F}_{GHT}(a^{i+1}_{r_h-1})\right)}$$

$$+ \sum_{\substack{h\in \bar{R}_{si}\\ r_h\in C_3}} \frac{\left(F_g(a^{i+1}_{r_h}) - F_g(b^{i+1}_{r_h})\right)\cdot\left(\widehat{k}_i(a^{i+1}_{r_h}) - \widehat{k}_i(a^{i+1}_{r_{(h-1)}})\right)}{N\cdot\left(\widehat{F}_{GHT}(a^{i+1}_{r_h}) - \widehat{F}_{GHT}(b^{i+1}_{r_h})\right)} =$$

$$+ \sum_{\substack{h=1\\ r_h\in C_1}}^{m_i} \frac{\left(F_g(a^{i+1}_{r_h}) - F_g(a^{i+1}_{r_h-1})\right)\cdot\left(\widehat{k}_i(a^{i+1}_{r_h}) - \widehat{k}_i(a^{i+1}_{r_{(h-1)}})\right)}{N\cdot\left(\widehat{F}_{GHT}(a^{i+1}_{r_h}) - \widehat{F}_{GHT}(a^{i+1}_{r_h-1})\right)}$$

$$+ \sum_{\substack{h=1\\ r_h\in C_3}}^{m_i} \frac{\left(F_g(a^{i+1}_{r_h}) - F_g(b^{i+1}_{r_h})\right)\cdot\left(\widehat{k}_i(a^{i+1}_{r_h}) - \widehat{k}_i(a^{i+1}_{r_{(h-1)}})\right)}{N\cdot\left(\widehat{F}_{GHT}(a^{i+1}_{r_h}) - \widehat{F}_{GHT}(b^{i+1}_{r_h})\right)}. \tag{3.32}$$

As in the Case 1) with $R_{s(i+1)} \neq \emptyset$ and $\bar{R}_{s(i+1)} \neq \emptyset$, the value $\widehat{F}_{YO}(z) - \widehat{F}_{YO}(t) \geq 0$.

If we assume that $\bar{R}_{s(i+1)} = \emptyset$, the sets C_3 and C_4 are empty. For $h \in \bar{R}_{si}$, then $r_h \in C_1$ and condition (3.27) is satisfied. From (3.19) and (3.27) it is easy to see that $\widehat{F}_{YO}(z)$ is given by (3.30) with the sum based on null C_3 and $T_s(z) = 0$.

Similarly, because $C_3 = \emptyset$; and the conditions(3.31) and (3.27) are satisfied, by (3.21) the value $\widehat{F}_{YO}(t)$ is given by (3.32) with the sum based on null C_3 . Thus, it is clear that $\widehat{F}_{YO}(z) - \widehat{F}_{YO}(t) \geq 0$.

Case 3) $\bar{R}_{si} = \emptyset$

In this case, in a similar way that in the previous case, for all $h = 1, \ldots, m_1$; the condition (3.31) is satisfied and consequently $r_h \in C_1$ and $C_3 = \emptyset$.

By (3.31) and (3.19), we have:

$$\widehat{F}_{YO}(t) = \sum_{\substack{h=1\\ r_h\in C_1}}^{m_i} \frac{\left(F_g(a^{i+1}_{r_h}) - F_g(a^{i+1}_{r_h-1})\right)\cdot\left(\widehat{k}_i(a^{i+1}_{r_h}) - \widehat{k}_i(a^{i+1}_{r_{(h-1)}})\right)}{N\cdot\left(\widehat{F}_{GHT}(a^{i+1}_{r_h}) - \widehat{F}_{GHT}(a^{i+1}_{r_h-1})\right)}.$$

If we assume that $R_{s(i+1)} \neq \emptyset$, the value $\widehat{F}_{YO}(z)$ is given by (3.30) with the sum based on null C_3 while if we

assume that $R_{s(i+1)} = \emptyset$ the value $\widehat{F}_{YO}(z)$ is given by (3.30) with the sum based on null C_3 and $T_s(z) = 0$. In any case, $\widehat{F}_{YO}(z) - \widehat{F}_{YO}(t) \geq 0$.

Definitely, in all cases, $\widehat{F}_{YO}(z) - \widehat{F}_{YO}(t) \geq 0$ if we consider $t \leq z$ with $y_{[i]} \leq t < y_{[i+1]}$ and $y_{[i+1]} \leq z < y_{[i+2]}$; $i = 1, \ldots, l-2$.

For $t \leq z$ with $y_{[i]} \leq t < y_{[i+1]}$ and $y_{[q]} \leq z < y_{[q+1]}$; $i = 1, \ldots, l-2$ and $q = 3, \ldots, l$ with $q > i+1$, by the previous cases, it is clear that:

$$\widehat{F}_{YO}(t) \leq \widehat{F}_{YO}(y_{[i+1]}) \leq \widehat{F}_{YO}(y_{[i+2]}) \leq \cdots \leq \widehat{F}_{YO}(y_{[q]}) \leq \widehat{F}_{YO}(z)$$

and the nondecreasing monotony of $\widehat{F}_{YO}(t)$ is proved.

Note

The estimator $\widehat{F}_{YO}(t)$ does not satisfy, in general, the condition $\lim_{t \to +\infty} \widehat{F}_{YO}(t) = 1$, but this condition is not strictly necessary as long as the following condition is satisfied

$$\text{máx}\{\widehat{F}_{YO}(y_i) : i \in s\} \geq \alpha. \tag{3.33}$$

Thus, we can define the following quantile estimator:

$$\widehat{Q}_{YO}(\alpha) = \text{ínf}\{t : \widehat{F}_{YO}(t) \geq \alpha\} = \widehat{F}_{YO}^{-1}(\alpha). \tag{3.34}$$

4. Variance estimation with resampling method

In this section we employ resampling techniques for the variance estimation of the quantile estimators proposed in Section 3 and the development of confidence intervals for quantiles associated with the calibration estimators proposed, because it is possible that a mathematic expression for their variance could be not establish due to the complexity of the estimators proposed (they are not linear functions of the data). More specifically, from a practical viewpoint we have considered the use of bootstrap techniques by their applicability in many cases and under different conditions.

Initially, the bootstrap method was developed by Efron (1979) under assumptions of an infinity population

with unknown distribution and the data is independently and identically distributed. Due to the popularity of this technique, the classical framework has been adjusted for survey sampling and incorporate the sampling design in several studies (Gross (1980), Booth et al. (1994), Chao & Lo (1994), Bickel & Freedman (1984), Chao & Lo (1994), Antal & Tillé (2011) and Antal & Tillé (2014)). Thus, Gross (1980), Booth et al. (1994) and Bickel & Freedman (1984) developed bootstrap methods where artificial populations are created from the sample by repeating its units and bootstrap samples are selected with the original sampling design from the artificial population. On the other hand, Antal & Tillé (2011) and Antal & Tillé (2014) consider direct bootstrap techniques where the bootstrap samples are obtained by units directly selected from the original sample under a completely different sampling scheme from the one which generated the original sample. In this study, we consider the frameworks proposed in Booth et al. (1994), Antal & Tillé (2011) and Antal & Tillé (2014).

Given a generic quantile estimator $\widehat{Q}_y(\alpha)$, following Booth et al. (1994), if $N = n \cdot c + a$ with $0 < a < n$, the artificial population is obtained by repeating c times the initial sample s and selecting by simple random sampling without replacement an additional sample of size a from the original sample s. The artificial population U_B is formed with this sample and the c replicates of s. Thus, let U_B^j with $j = 1, \ldots, M$ be M independent artificial populations obtained from s, for each pseudo population U_B^j we select K bootstrap samples $s_1^j, \ldots, s_K^j$ with sample size n. Next, following (Chauvet (2007)), we compute the bootstrap estimates $\widehat{Q}_y^*(\alpha)_h^j$ with the sample s_h^j for the population U_B^j and we consider:

$$\widehat{V}_j = \frac{1}{K-1} \sum_{h=1}^{K} (\widehat{Q}_y^*(\alpha)_h^j - \widehat{Q}_y^*(\alpha)^j)^2 \tag{3.35}$$

where

$$\widehat{Q}_y^*(\alpha)^j = \frac{1}{K} \sum_{h=1}^{K} \widehat{Q}_y^*(\alpha)_h^j.$$

Now, the variance estimation for the estimator $\widehat{Q}_y(\alpha)$ is given by:

$$\widehat{V}(\widehat{Q}_y(\alpha)) = \frac{1}{M} \sum_{j=1}^{M} .\widehat{V}_j. \tag{3.36}$$

Recently, Antal & Tillé (2011) and Antal & Tillé (2014) have proposed direct boostrap methods where it is not necessary to obtain an artificial population, since the boostrap samples are obtained from the original sample by means of a sampling design different from the original sampling design considered. Thus, when the original sample s is obtained with simple random sampling without replacement, Antal & Tillé (2011)

has proposed a mixture sampling design where two samples are selected from s, one obtained by simple random sampling without replacement and the other sample obtained with one-one sampling design (a sampling design defined by the authors for resampling). Similarly, when s is obtained by simple random sampling without replacement, Antal & Tillé (2014) has proposed a mixture sampling scheme where the first sample is obtained with Bernoulli design while the second one is obtained with another sampling designed for resampling called double half sampling design by the authors. For more details on the two direct boostrap methods included in this study, see Antal & Tillé (2011) and Antal & Tillé (2014).

In both frameworks, given a generic quantile estimator $\widehat{Q}_y(\alpha)$, for the original sample s, we select M bootstrap samples $s_1^*, \ldots, s_M^*$ according to the sampling schemes of Antal & Tillé (2011) and Antal & Tillé (2014) respectively. The variance bootstrap estimation for the estimator $\widehat{Q}_y(\alpha)$ is given by:

$$\widehat{V}(\widehat{Q}_y(\alpha)) = \frac{1}{M}\sum_{j=1}^{M}(\widehat{Q}_y(\alpha)_j^* - \bar{Q}_y(\alpha)^*)^2 \tag{3.37}$$

where $\widehat{Q}_y(\alpha)_j^*$ is the bootstrap estimator computed on the bootstrap sample s_j^* and

$$\bar{Q}_y(\alpha)^* = \frac{1}{M}\sum_{j=1}^{M}\widehat{Q}_y(\alpha)_j^*.$$

Finally, for a quantile estimator $\widehat{Q}_y(\alpha)$ with a variance estimation $\widehat{V}(\widehat{Q}_y(\alpha))$ obtained with a bootstrap method, we consider the $1-\alpha$ level confidence interval based on the approximation by a standard normal distribution:

$$\left[\widehat{Q}_y(\alpha) - z_{1-\alpha/2}\cdot\widehat{V}(\widehat{Q}_y(\alpha)), \widehat{Q}_y(\alpha) + z_{1-\alpha/2}\cdot\widehat{V}(\widehat{Q}_y(\alpha))\right] \tag{3.38}$$

where z_α denotes the α quantile of the standard normal distribution. For the three proposed bootstrap methods included in this study, we can obtain with this procedure the respective confident interval.

5. Application of the optimal quantile estimators in poverty measures estimation

For governments it is of high interest the estimation of poverty and wage inequality. Inequality and life condition indicators and many social indicators related to the measurement of poverty are based upon quantiles. Among the poverty measures commonly used by institutions in their reports on poverty, we can find the poverty line and the Head Count Index. For instance, Eurostat establishes poverty line as 60 percent of the median of the equivalized net income. Thus, the poverty line is defined as a threshold that divides

the population into poor and nonpoor that depends on the median value. The Head Count Index (HCI) can be calculated as the proportion of persons (or households) with an equivalised disposable income below the poverty line. On the other hand, some measures for wage inequality employed in several studies are based on percentiles ratios like 50th/5th and 50th/25th (Dickens & Manning (2004)); 50th/10th (Nickell (2004),Machin et al. (2003), Metcalf (2008), Burtless (1999)); 95th/50th (Machin et al. (2003), Burtless (1999)) and 90th/10th; 95th/20th; and 80th/20th (Jones & Weinberg (2000)). In this study, we focus on the estimation of the poverty measures based on percentile ratios.

Thus, for a finite population $U = \{1, \ldots, N\}$ with distribution function $F_y(t)$ given by (4.1), the percentile ratio $R(\alpha_1, \alpha_2)$ is defined as follows:

$$R(\alpha_1, \alpha_2) = \frac{Q_y(\alpha_1)}{Q_y(\alpha_2)} = \frac{F_y^{-1}(\alpha_1)}{F_y^{-1}(\alpha_2)}$$

and evidently, it can be estimated with the quantile estimator $\tilde{Q}_{YO}(\alpha)$ as follow:

$$\tilde{R}_{YO}(\alpha_1, \alpha_2) = \frac{\tilde{Q}_{YO}(\alpha_1)}{\tilde{Q}_{Y1}(\alpha_2)}.$$

Obviously, the variance estimation of a percentile ratio estimator present similar drawbacks to the estimation of variance for quantile estimator and consequently, we can compute the estimation of variance for $\tilde{R}_{YO}(\alpha_1, \alpha_2)$ and confidence intervals for $\widehat{R}_{YO}(\alpha_1, \alpha_2)$ with the resampling techniques described in the previous section.

6. Simulation study

This section provides numerical comparisons for some poverty measure estimators proposed in Sections 3 and 5. In two simulation studies the proposed estimators are compared with the corresponding poverty measures estimators derived from previous estimators of the distribution function: the Horvitz-Thompson estimator $\widehat{F}_{YHT}(t)$, the difference estimator $\widehat{F}_{YD}(t)$ (see Rao et al. (1990)), the ratio estimator $\widehat{F}_{YR}(t)$ (see Rao et al. (1990)), the Chambers–Dunstan estimator $\widehat{F}_{YCD}(t)$ (see Chambers & Dunstan (1986)) and the Rao,Kovar and Mantel estimator (see Rao et al. (1990)) $\widehat{F}_{YRKM}(t)$. Additionally, we have included the quantile estimator and the estimator of poverty measures derived from the calibrated estimator $\widehat{F}_{yc}(t)$ of the distribution function proposed in Rueda et al. (2007a), with auxiliary vector $\mathbf{t_g} = (Q_g(0{,}25), Q_g(0{,}5), Q_g(0{,}75))$ and we denoted by $\widehat{F}_{YQUAR}(t)$ the corresponding calibration estimator. Some of these estimators $\widehat{F}_y(t)$ included in the simulation study are not monotonically nondecreasing functions; for these estimators we have considered

a general procedure described in Rao et al. (1990) to obtain a monotonous nondecreasing version of the estimator $\tilde{F}_y(t)$.

For both simulation studies, the estimation of the variance provided by the bootstrap methods included in Section 4 is also analyzed. All simulations included in this section have been developed with new code programmed in R.

In the first study we consider real data from the region of Cantabria of the 2008 Spanish living conditions survey carried out by the Instituto Nacional de Estadística (INE) of Spain. The survey data collected are considered as a population with size $N = 377$ and samples are selected from it. In this study we obtain estimation of the poverty threshold L, where L is calculated following the criteria recommended by Eurostat, that is, the threshold L is set at 60 % of the median of the equivalised net income (the study variable). We considered the attribute "Home with own computer" as the auxiliary variable. We selected $W = 1000$ samples with several sample sizes, n, under SRSWOR and for each estimator included in the simulation study, we computed estimates of the poverty threshold L. The performance of each estimator is measured by the relative bias (RB) and the relative efficiency (RE), given respectively by

$$\mathrm{RB}(\widehat{L}) = \frac{1}{W}\sum_{w=1}^{W}\frac{(\widehat{L}_w - L)}{L} \tag{3.39}$$

$$\mathrm{RE}(\widehat{L}) = \frac{\sum_{w=1}^{W}\left[\widehat{L}_w - L\right]^2}{\sum_{w=1}^{W}\left[\left(\widehat{L}_{HT}\right)_w - L\right]^2}, \tag{3.40}$$

where w indexes the wth simulation run; $\widehat{L}$ is a poverty threshold estimator and $\widehat{L}_{HT}$ is the poverty threshold estimator based in the Horvitz-Thompson $\widehat{F}_{YHT}(t)$ estimator.

From every simulation sample, 1000 bootstrap samples were selected using the three bootstrap methods considered in Section 4, for the variance estimation and confidence intervals. We computed the following measures: the coverage probability (CP), the lower (L) and the upper (U) tail error rates of the 95 % confidence intervals, in percentage and the average length (AL) of the confidence intervals for each estimator and each bootstrap method, except for the Chambers-Dunstan estimator whose results are only obtained with the Booth method, since this estimator needs the whole population for its calculation and the techniques described do not obtain the whole artificial population. Results from this simulation study are presented in Table A3.1 and Table A3.2.

Tabla A3.1: RB and RE for several sample sizes of the estimators compared. SRSWOR from the 2008 SPANISH LIVING CONDITIONS SURVEY.

	RB	RE	RB	RE	RB	RE	RB	RE
Estimator	$n = 50$		$n = 60$		$n = 70$		$n = 80$	
$\widehat{L}_{HT}$	-0.0159	1.0000	-0.0103	1.0000	-0.0144	1.0000	-0.0118	1.0000
$\widehat{L}_{CD}$	0.0452	1.2561	0.0408	1.1163	0.0332	1.0551	0.0324	1.0732
$\tilde{L}_{d}$	-0.0038	0.9246	-0.0010	0.8975	-0.0056	0.8796	-0.0044	0.8873
$\tilde{L}_{r}$	-0.0069	1.7426	-0.0072	1.5946	-0.0057	1.4706	-0.00603	1.6292
$\tilde{L}_{RKM}$	-0.0043	0.9456	-0.0014	0.8979	-0.0058	0.8942	-0.0041	0.8971
$\widehat{L}_{YQUAR}$	-0.0159	1.0000	-0.0103	1.0000	-0.0144	1.0000	-0.0118	1.0000
$\widehat{L}_{YCO}$	-0.0039	0.9165	-0.0005	0.8974	-0.0056	0.8730	-0.0043	0.8849

The results derived from this simulation study gave values for RB within a reasonable range. The proposed estimator significantly improve the results of the calibrated estimator $\widehat{L}_{YQUAR}$. With respect to efficiency, the best estimator for all sample sizes is $\widehat{L}_{YCO}$ whereas the usual calibrated estimator have an efficiency similar to $\tilde{L}_{HT}$.

With respect to the variance estimation, all estimators provide high coverages, with values very close to 99 % in the three resampling methods considered. For the resampling methods proposed in Booth et al. (1994) and Antal & Tillé (2014), the proposed estimators present the best average length (AL) results for some sample sizes, whereas with the method proposed in Antal & Tillé (2011), the proposed estimators present the best results for all the sample sizes, with the exception of size $n = 60$.

For the second simulation study, we consider real data from the region of Andalusia of 2016 Spanish living conditions survey carried out by the Instituto Nacional de Estadística (INE) of Spain. The survey data collected are considered as a population with size $N = 1442$ and samples are selected from it. The study variable y is the equivalised net income and the auxiliary variables included are the attribute "Can the home afford to go on vacation away from home, at least one week a year?", the attribute "Home with own computer" and the attribute "Home with own washing machine" as the auxiliary variables. Again, we selected $W = 1000$ samples with several sample sizes, $n = 75$, $n = 95$, $n = 115$ and $n = 135$, under SRSWOR and for each estimator included in the simulation study, we computed estimates of $R(\alpha_1, \alpha_2)$ for 95th/50th. The performance of each estimator is measured by the values RB and RE, given by

$$\mathrm{RB}(\widehat{R}(\alpha_1, \alpha_2)) = \frac{1}{W} \sum_{w=1}^{W} \frac{\left(\widehat{R}(\alpha_1, \alpha_2)\right)_w - R(\alpha_1, \alpha_2)}{R(\alpha_1, \alpha_2)} \tag{3.41}$$

Tabla A3.2: AL, CP %,L % and U % for several sample sizes and several resampling methods of the estimators compared.SRSWOR from the 2008 SPANISH LIVING CONDITIONS SURVEY.

	AL	CP %	L %	U %	AL	CP %	L %	U %	AL	CP %	L %	U %	AL	CP %	L %	U %
									Booth et al., (1994)							
Estimator	$n = 50$				$n = 60$				$n = 70$				$n = 80$			
$\widehat{L}_{HT}$	7355	99.9	0.0	0.1	4513	100.0	0.0	0	3554	100.0	0	0.0	3199	100.0	0.0	0
$\widehat{L}_{CD}$	7323	99.9	0.1	0.0	3670	99.8	0.2	0	4451	100.0	0	0.0	4082	99.1	0.1	0
$\tilde{L}_d$	7383	100.0	0.0	0.0	4164	100.0	0.0	0	3501	100.0	0	0.0	3103	100.0	0.0	0
$\tilde{L}_r$	7397	99.8	0.2	0.0	6001	100.0	0.0	0	3834	99.9	0	0.1	3423	99.8	0.2	0
$\tilde{L}_{RKM}$	7096	100.0	0.0	0.0	4171	100.0	0.0	0	3418	100.0	0	0.0	3041	100	0.0	0
$\widehat{L}_{YQUAR}$	7355	99.9	0.0	0.1	4513	100.0	0.0	0	3454	100.0	0	0.0	3115	100	0.0	0
$\widehat{L}_{YCO}$	7301	100.0	0.0	0.0	4166	100.0	0.0	0	3507	100.0	0	0.0	3016	100	0.0	0
									Antal, E., Tillé, Y. (2014)							
$\widehat{L}_{HT}$	5198	98.9	0.1	1.0	3970	98.4	0.2	1.4	3507	98.0	0.4	1.6	3072	98.1	0.4	1.5
$\tilde{L}_d$	5204	98.9	0.4	0.7	3477	98.7	0.4	0.9	2914	98.3	0.6	1.1	2504	98.4	0.5	1.1
$\tilde{L}_r$	5518	99.0	0.2	0.8	4108	99.4	0.0	0.6	360	98.8	0.1	1.1	316898.7	0.2	1.1	
$\tilde{L}_{RKM}$	4659	99.1	0.2	0.7	4224	98.9	0.3	0.8	3001	98.0	0.7	1.3	2671	98.1	0.6	1.3
$\widehat{L}_{YQUAR}$	5198	98.9	0.1	1.0	3970	98.4	0.2	1.4	3507	98.0	0.4	1.6	3044	98.3	0.3	1.4
$\widehat{L}_{YCO}$	4754	98.7	0.5	0.8	4235	98.5	0.5	1.0	2938	98.1	0.6	1.3	2448	98.4	0.4	1.2
									Antal, E., Tillé, Y. (2011).							
$\widehat{L}_{HT}$	4417	98.9	0.1	1.0	4627	98.0	0.3	1.7	2488	98.4	0.2	1.4	2279	98.5	0.2	1.3
$\tilde{L}_d$	3219	99.1	0.1	0.8	3855	98.6	0.5	0.9	2511	98.9	0.5	0.6	2281	99.2	0.4	0.4
$\tilde{L}_r$	4329	99.3	0.2	0.5	4957	99.5	0.1	0.4	4177	98.7	0.3	1.0	3837	98.9	0.3	0.8
$\tilde{L}_{RKM}$	3846	98.9	0.0	1.1	4088	98.7	0.6	0.7	2724	98.8	0.5	0.7	2435	99.1	0.5	0.4
$\widehat{L}_{YQUAR}$	4417	98.9	0.1	1.0	4627	98.0	0.3	1.7	2488	98.4	0.2	1.4	2173	98.6	0.2	1.2
$\widehat{L}_{YCO}$	3059	98.9	0.2	0.9	4188	98.2	0.9	0.9	2309	98.7	0.5	0.8	2018	99.3	0.3	0.4

$$\mathrm{RE}(\widehat{R}(\alpha_1,\alpha_2)) = \frac{\sum_{w=1}^{W}\left[\left(\widehat{R}(\alpha_1,\alpha_2)\right)_w - R(\alpha_1,\alpha_2)\right]^2}{\sum_{w=1}^{W}\left[\left(\widehat{R}_{HT}(\alpha_1,\alpha_2)\right)_w - R(\alpha_1,\alpha_2)\right]^2}, \tag{3.42}$$

where $\widehat{R}(\alpha_1,\alpha_2)$ is a percentile ratio estimator and $\widehat{R}_{HT}(\alpha_1,\alpha_2)$ is the percentile ratio estimator based in the Horvitz-Thompson $\widehat{F}_{YHT}(t)$ estimator .

For the variance estimation and confidence intervals, we computed the coverage probability (CP), the lower (L) and the upper (U) tail error rates of the 95 % confidence intervals, in percentage and the average length (AL) of the confidence intervals for each percentile ratio estimator and each bootstrap method.

Concerning the variance estimation and confidence intervals, we used 1,000 bootstrap replications from each initial sample with all bootstrap methods included in the study to compute CP, L, U and AL of the 95 % confidence intervals for each percentile ratio considered. Result from this simulation study are presented in Table A3.3 and Table A3.4.

Tabla A3.3: RB and RE for several sample sizes of the estimators compared. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY.

	RB	RE %	RB	RE	RB	RE	RB	RE
Estimator	$n = 75$		$n = 95$		$n = 115$		$n = 135$	
$\widehat{R}_{HT}$	0.0392	1	0.0358	1	0.0285	1	0.0234	1
$\widehat{R}_{CD}$	-0.1224	1.0838	-0.1152	1.0771	-0.1219	1.3093	-0.125	1.4158
$\tilde{R}_d$	0.0407	1.0394	0.0353	1.0467	0.0269	1.0673	0.0223	1.0689
$\tilde{R}_r$	0.0461	1.3042	0.0406	4.0376	0.018	4.4338	0.0161	3.2556
$\tilde{R}_{RKM}$	0.0333	1.0384	0.033	1.0475	0.0263	1.0325	0.0208	1.0266
$\widehat{R}_{YQUAR}$	0.0251	0.948	0.0208	0.9276	0.0183	0.9635	0.0143	0.9623
$\widehat{R}_{YCO}$	0.0174	0.921	0.0157	0.8871	0.0164	0.9458	0.0126	0.9390

These tables show:

- The percentile ratio estimator based on the Chambers–Dunstan estimator has a serious problem of bias. This is expected because the estimator $\widehat{F}_{YCD}(t)$ is biased when the relation between y and x is not linear. We found no evidence of any significant bias for the other estimators considered.
- In terms of efficiency the best overall performance is achieved by our proposed calibration estimator. This estimator performs remarkably better than the other estimators.
- The three methods of estimating the variances provide intervals with coverage below the nominal coverage. Although there is not much difference between the methods it seems that the first method (Booth et al. (1994)) provides narrower intervals.

Tabla A3.4: AL, CP %,L % and U % for several sample sizes and several resampling methods of the estimators compared. SRSWOR from the 2016 SPANISH LIVING CONDITIONS SURVEY.

	AL	CP %	L %	U %	AL	CP %	L %	U %	AL	CP %	L %	U %	AL	CP %	L %	U %
									Booth et al., (1994)							
Estimator	$n = 75$				$n = 95$				$n = 115$				$n = 135$			
$\widehat{R}_{HT}$	3.7901	87.7	3.8	8.5	0.8089	87.4	5.1	7.5	1.7979	86.9	5.6	7.5	0.7069	87.3	5	7.7
$\widehat{R}_{CD}$	3.5177	66	1	33	1.4043	63.3	1.2	35.5	0.2989	60.9	0.7	38.4	0.5982	56.8	0.5	42.7
$\tilde{R}_{d}$	3.8003	85.9	4.4	9.7	1.1235	87.9	5	7.1	1.7837	86.7	5.3	8	0.6935	86.5	5.4	8.1
$\tilde{R}_{r}$	3.7722	82.7	7.7	9.6	1.8362	84.4	6.8	8.8	1.0592	83.4	6	10.6	0.991	83.6	5.3	11.1
$\tilde{R}_{RKM}$	3.8459	87.5	3.5	9	1.2744	87.9	4.5	7.6	1.5897	86	5.2	8.8	0.6935	86.2	5.3	8.5
$\widehat{R}_{YQUAR}$	3.8042	86.4	3.8	9.8	1.2659	88	4.2	7.8	1.7837	85.9	4.8	9.3	0.6935	87.3	4.4	8.3
$\widehat{R}_{YCO}$	3.6703	85.6	4	10.4	1.2659	87.8	3.8	8.4	1.7837	85.4	5.3	9.3	0.6935	88.2	3.6	8.2
									Antal, E., Tillé, Y. (2014)							
$\widehat{R}_{HT}$	3.8038	84.7	6	9.3	1.1358	82.6	8.6	8.8	1.7578	85.2	6.7	8.1	1.3928	82.8	8.8	8.4
$\tilde{R}_{d}$	3.9	85.3	5.2	9.5	1.1534	82.5	8.7	8.8	1.4795	84.5	6.5	9	1.2933	81.3	8.8	9.9
$\tilde{R}_{r}$	4.4885	78.9	7.6	13.5	1.4523	80.8	6.7	12.5	2.1691	80.9	6.8	12.3	2.2027	77.9	8.9	13.2
$\tilde{R}_{RKM}$	3.8207	84.8	5.2	10	1.2017	82.6	8.7	8.7	1.4795	84.6	7.2	8.2	1.4643	82.2	7.8	10
$\widehat{R}_{YQUAR}$	3.4395	83.4	5.4	11.2	1.1509	84.3	7.1	8.6	1.527	83.7	7.1	9.2	1.2797	82.9	7.3	9.8
$\widehat{R}_{YCO}$	3.4395	84.3	4.5	11.2	1.1336	84.8	6.5	8.7	1.8641	82.7	7.3	10	1.4264	83.1	7	9.9
									Antal, E., Tillé, Y. (2011)							
$\widehat{R}_{HT}$	3.8276	83.4	6.3	10.3	0.5086	83.3	7.8	8.9	1.3731	85.4	6.6	8	1.1126	82.9	8.9	8.2
$\tilde{R}_{d}$	4.7222	83	5.5	11.5	1.0117	83.7	8.2	8.1	1.3943	84.4	7	8.6	1.0308	81.3	9.1	9.6
$\tilde{R}_{r}$	3.6417	78.9	8.7	12.4	3.5245	79.4	8	12.6	1.0688	77.8	9.3	12.9	1.2873	79.5	8.8	11.7
$\tilde{R}_{RKM}$	4.6509	82.6	5.8	11.6	1.0117	83.4	7.6	9	1.3963	85.8	5.5	8.7	1.0308	81.9	8.8	9.3
$\widehat{R}_{YQUAR}$	4.7222	82.2	5.7	12.1	1.5538	84.7	6.4	8.9	1.3963	85.2	5.4	9.4	0.9372	81.6	8.4	10
$\widehat{R}_{YCO}$	3.8444	82.5	5.6	11.9	2.0123	83.8	6.5	9.7	1.3281	85.2	4.9	9.9	1.0308	82.6	7.2	10.2

To sum up, these simulation show how the use of the auxiliary information by the proposed estimators can reduce the error of the usual direct and indirect estimators. Overall, the proposed estimators $\widehat{L}_{YCO}$ and $\widehat{R}_{YCO}$, appear to be good estimators.

It is also remarkable that the bootstrap method tends to overestimate the variance for the poverty threshold L, whenever that bootstrap variance is smaller than the variance in the case of the 95th / 50th percentile ratio. This is not surprising since the bootstrap technique for nonlinear parameters does not provide unbiased estimators of the variances and rescaling may be necessary to achieve exact unbiasedness (Wolter, 2007). This same problem appears in the results obtained in the simulations performed by Antal & Tillé (2011, 2014). In these simulations the bootstrap variance estimators are also strongly biased when applied to quantiles and poverty measures.

7. Conclusions

In this paper we investigate the optimum estimation of the quantiles in the sense of minimum variance. We start from the calibration estimator proposed by Rueda et al. (2007b) and transform the problem of minimizing the variance of this quantile estimator into a problem of minimizing the variance of the estimator of the associated distribution function. Besides, we obtain an optimal estimator of the distribution function $\widehat{F}_{YO}(t)$ that is a genuine function of distribution and therefore does not need the procedure to satisfy the non-decreasing monotony as the $\tilde{Q}_d$ and $\tilde{Q}_{RKM}$ estimators. The simulation studies indicate that calibrated estimators proposed in the present work are also a suitable option for the estimation of measures for wage inequality based on percentiles ratios and poverty lines, but the simulation also shows that the bootstrap estimators for the poverty measures does not provide unbiased estimators of the variances and rescaling may be necessary to achieve exact unbiasedness.

Acknowledgments

This study was partially supported by Ministerio de Educación y Ciencia (grant MTM2015-63609-R, Spain)

References

[1] Antal, E., & Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106(494), 534-543.

[2] Antal, E., & Tillé, Y. (2014). A new resampling method for sampling designs without replacement: the doubled half bootstrap. *Computational Statistics*, 29(5), 1345–1363.

[3] Arcos, A., Martínez, S., Rueda, M., & Martínez, H. (2017). Distribution function estimates from dual frame context, *Journal of Computational and Applied Mathematics*, 318, 242-252.

[4] Arcos, A., Rueda, M., & Muñoz, J.F. (2007). An improved class of estimators of a finite population quantile in sample surveys *Applied Mathematics Letters*, 20(3), 312-315.

[5] Bickel, P. J., & Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12(2), 470–482.

[6] Bogin, B., & Sullivan, T. (1986). Socioeconomic status, sex, age, and ethnicity as determinants of body fat distribution for Guatemalan children. *American Journal of Physical Anthropology*, 69(4), 527-535.

[7] Booth, J. G., R. W. Butler, R. W., & Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89(428), 1282–1289.

[8] Burtless, G. (1999). Effects of growing wage disparities and changing family composition on the US income distribution. *European Economic Review*, 43(4-6), 853-865.

[9] Chambers, R. L., & Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597-604.

[10] Chao, M. T., & Lo, S.H. (1994). Maximum likelihood summary and the bootstrap method in structured finite populations. *Statistica Sinica*, 4(2), 389–406.

[11] Chauvet, G. (2007) Méthodes de bootstrap en population finie. PhD thesis, Université Rennes 2.

[12] Chen, J., & Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 385-406.

[13] Chen, J., & Wu, C., 2002. Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 1223-1129.

[14] Decker, R., Haltiwanger, J., Jarmin, R., & Miranda, J. (2014). The role of entrepreneurship in US job creation and economic dynamism. *Journal of Economic Perspectives*, 28(3), 3-24.

[15] Deville,J.C., & Särndal, C.E.(1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87(418), 376-382.

[16] Dickens, R., & Manning, A. (2004). Has the national minimum wage reduced UK wage inequality?. *Journal of the Royal Statistical Society: Series A*, 167(4), 613-626.

[17] Dorfman, A. H., & Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, 21(3), 1452-1475.

[18] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.

[19] Gross, S. (1980). Median estimation in sample surveys. In Proceedings of the Section on Survey Research Methods, *American Statistical Association*, 181–184.

[20] Harms, T., & Duchesne, P. (2006). On calibration estimation for quantiles, *Survey Methodology*, 32, 37-52.

[21] INE (2015). Continuous Register Statistics.
https://www.ine.es

[22] Jones Jr, A. F., & Weinberg, D. H. (2000). The changing shape of the nation's income distribution, 1947-1998. Current Population Reports P60-204. Washington, DC:U.S. Census Bureau.
http://www.census.gov/ftp/pub/hhes/www/p60204.html

[23] Kuk, A. Y., & Mak, T. K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(2), 261-269.

[24] Machin, S., Manning, A., & Rahman, L. (2003). Where the minimum wage bites hard: introduction of minimum wages to a low wage sector. *Journal of the European Economic Association*, 1(1), 154-180.

[25] Martínez, S., Rueda, M., Arcos, A., & Martínez, H. (2010). Optimum calibration points estimating distribution functions. *Journal of Computational and Applied Mathematics*, 233(9), 2265-2277.

[26] Martínez, S., Rueda, M., Arcos, A., Martínez, H., & Sánchez-Borrego, I. (2011). Post-stratified calibration method for estimating quantiles. *Computational Statistics and Data Analysis*, 55(1), 838-851.

[27] Martínez, S., Rueda, M., Arcos, A., Martínez, H., & Muñoz, J. F. (2012). On determining the calibration equations to construct model-calibration estimators of the distribution function. *Revista Matemática Complutense*, 25(1), 87-95.

[28] Martínez, S., Rueda, M., Martínez, H., & Arcos, A. (2015). Determining P optimum calibration points to construct calibration estimators of the distribution function. *Journal of Computational and Applied Mathematics*, 275, 281-293.

[29] Martínez, S., Rueda, M., Martínez, H., & Arcos, A. (2017). Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function. *Journal of Computational and Applied Mathematics*, 318, 444-459.

[30] Mayor-Gallego, J. A., Moreno-Rebollo, J. L., & Jiménez-Gamero, M. D. (2019). Estimation of the finite population distribution function using a global penalized calibration method. *AStA Advances in Statistical Analysis*, 103(1), 1-35.

[31] Metcalf, D. (2008). Why has the British national minimum wage had little or no impact on employment?. *Journal of Industrial Relations*, 50(3), 489-512.

[32] Morales, D., Rueda, M., & Esteban, D. (2018). Model-assisted estimation of small area poverty measures: An application within the Valencia Region in Spain. *Social Indicators Research*, 138, 873–900.

[33] Nickell, S. (2004). Poverty and worklessness in Britain. *The Economic Journal*, 114(494), C1-C25.

[34] Rao, J. N. K. and Kovar, J. G. & Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information, *Biometrika*, 77(2), 365-375.

[35] Rueda, M., Martinez, S., Martinez, H., & Arcos, A. (2006). Mean estimation with calibration techniques in presence of missing data. *Computational Statistics and Data Analysis*, 50(11), 3263-3277.

[36] Rueda, M., Martínez, S., Martínez, H., & Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137(2), 435-448.

[37] Rueda, M., Martínez-Puertas, S., Martínez-Puertas, H., & Arcos, A. (2007). Calibration methods for estimating quantiles. *Metrika*, 66(3), 355-371.

[38] Rueda, M., Martinez, S., Arcos, A., & Munoz, J. F. (2009). Mean estimation under successive sampling with calibration estimators. *Communications in Statistics-Theory and Method*, 38(6), 808-827.

[39] Särndal C.E. (2007) The calibration approach in survey theory and practice. *Survey Metholodogy*, 33(2), 99–119.

[40] Singh, H. P., Singh, S., & Kozak, M. (2008). A family of estimators of finite-population distribution function using auxiliary information. *Acta applicandae mathematicae*, 104(2), 115-130.

[41] Tellez-Plaza, M., Navas-Acien, A., Crainiceanu, C. M., & Guallar, E. (2008). Cadmium exposure and hypertension in the 1999–2004 National Health and Nutrition Examination Survey (NHANES). *Environmental health perspectives*, 116(1), 51-56.

Apéndice A4

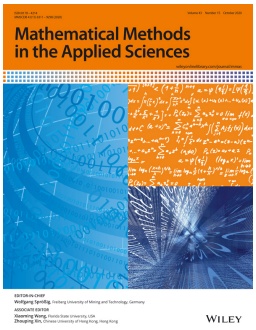
Reduction of optimal calibration dimension with a new optimal auxiliary vector for calibrated estimators of the distribution function

Martínez, Sergio; Rueda, María del Mar; Illescas, María Dolores (2022)

Reduction of optimal calibration dimension with a new optimal auxiliary vector for calibrated estimators of the distribution function.

Mathematical Methods in the Applied Sciences, Vol. 45, num. 17, pp. 10959–10981.

DOI: 10.1002/mma.8431



MATHEMATICS, APPLIED			
JCR Year	Impact factor	Rank	Quartile
2021	3.007	29/267	Q1

Abstract

The calibration method (Deville & Särndal, 1992) has been widely used to incorporate auxiliary information in the estimation of various parameters. Specifically, Rueda et al. (2007a) adapted this method to estimate the distribution function, although their proposal is computationally simple, its efficiency depends on the selection of an auxiliary vector of points. This work deals with the problem of selecting the calibration auxiliary vector that minimize the asymptotic variance of the calibration estimator of distribution function. The optimal dimension of the optimal auxiliary vector is reduced considerably with respect to previous studies(Martínez et al., 2017) so that with a smaller set of points the minimum of the asymptotic variance can be reached, which in turn allows to improve the efficiency of the estimates.

1. Introduction

In sample surveys, auxiliary population information is sometimes used in the estimation stage to increase the precision of the estimators of a mean or total population. Previous literature has investigated the use of auxiliary information to improve the estimation of a finite population mean, however, previous studies have considered to a lesser extent the development of efficient methods to estimate the distribution function and the finite population quantiles by incorporating the auxiliary information. The estimation of finite population distribution function is an important issue because the distribution function can be more useful than means and totals (Sedransk & Sedransk, 1979).Through the finite population distribution function, parameters such as population quantiles can be obtained. More specifically, in economics, many indicators used in the poverty analisys are based on quantiles, since they analyze variables with skewed distributions such as income, and in such cases the median is a more suitable location measure than the mean. Moreover, poverty studies incorporate the analysis of wage inequality and income distribution thorugh percentile ratios (Dickens & Manning, 2004; Nickell, 2004; Machin et al., 2003).

In the last decade, the well-known calibration estimation method to estimate the population total (Deville & Särndal, 1992) has been employed to develop new estimators which incorporates the auxiliary information available and it has become an important field of research in survey sampling (Rueda et al., 2007a; Estevao & Särndal, 2006; Singh, 2001; Devaud & Tillé, 2019; Rueda, 2019).

Previous works (Rueda et al., 2007a; Kovacevic, 1997; Harms & Duchesne, 2006) use different implementations of the calibration approach to obtain estimators of the distribution function and the quantiles.

Under a general superpopulation model Wu (2003) propose a model-calibrated estimators that is optimal under a chosen model with respect to the anticipated variance. Although Wu (2003) considers a general sampling design, its proposal does not produce an estimator with the properties of a genuine distribution function unless the weight system is obtained by using a point t_0 for any t value, which restricts the efficiency of the estimator to a neighborhood of t_0. Additionally, the proposal Wu (2003) requires the estimation of certain superpopulation parameters that depend on the study variable, which may restrict its applicability in some cases and also require additional conditions on the sampling design to maintain the asymptotic behavior of the proposed estimator Chen & Wu (2002).

Nonparametric regression (Breidt et al., 2007; Rueda et al., 2007a)), is also used for model-calibration estimation of the distribution function. Mayor-Gallego et al. (2019) propose a new estimator for the distribution function that integrates ideas from model calibration and penalized calibration. The method Rueda et al. (2007a) is computationally simple and it employs the calibration method by minimizing the chi-square distance subject to calibration equations that require the use of arbitrarily fixed values. One drawback of these estimators is that their efficiency depends on selected points. Under simple random sampling, the problem of optimal selection points in order to obtain the best estimation has been treated in previous works (Martínez et al., 2017, 2010, 2011, 2015). In fact, the work Martínez et al. (2017) obtained the optimal dimension and the optimal auxiliary vector for the estimator of the distribution function proposed in the work Rueda et al. (2007a) and although this proposal do not generate a unique weight system that is optimal for each point t, it produces an estimator that is computationally simple and is a genuine distribution function that can be used directly in the estimation of quantiles and poverty measures (Martínez et al., 2022).

In many situations, the optimal auxiliary vector has a very high dimension, which makes the calibration process difficult and can also affect the efficiency of the estimator. Performing calibration with a high dimensional auxiliary dataset can be several problems: the variance of the calibration estimator can be increases and the optimisation procedure may fail. Nascimento Silva & Skinner (1997) showed that if too many auxiliary variables are used, the bias of the calibrated estimator increases and can become nonnegligible compared to the variance (over-calibration). Recently Chauvet & Goga (2022) theoretically prove that over-calibration may deteriorate the efficiency of the estimates. Various procedures have been suggested for variable selection. Nascimento Silva & Skinner (1997) computed the mean squared error (MSE) for all possible subsets of quantitative auxiliary variables and then chose the one producing the smallest MSE. Later, Chambers & Clark (2008) used forward and stepwise selection based on the difference between the MSE of the prediction for two nested sets of variables. Alternatively, the least absolute shrinkage and selection operator (LASSO) (McConville et al., 2017) might be considered for selecting the best subsets. Once the best set of regressors has been selected, the calibration is performed on these variables alone.

Another approach to consider is that of penalised calibration (Guggemos & Tillé, 2010), which takes account of auxiliary information by attaching more or less importance according to its presumed explanatory power for the variable of interest. In a different way, Cardot et al. (2017) and Rota (2017) suggested applying principal component analysis for quantitative auxiliary variables in order to achieve a strong dimension reduction. These works are oriented to the estimation of linear parameters.

In this work, we intend to analyze whether it is possible to reduce the optimal dimension of the auxiliary vector proposed in the previous work Martínez et al. (2017). The remainder of the article is organized as follow. After introducing the problem of distribution function estimation in Section 2 with the method proposed in research work Rueda et al. (2007a) and the optimal auxiliary vector proposed in the previous work Martínez et al. (2017), in Section 3 we will analyze the conditions under which we can reduce the dimension of the optimal auxiliary vector. Then, Section 4 proposes a new calibration estimator based on the results of Section 3. Section 5 reports the results of an extensive simulation study run on a set of synthetic and real finite populations in which the performance of the proposed class of estimators is investigated for finite size samples. Section 6 provides some conclusions.

2. Calibration estimation of the distribution function and optimal auxiliary vector

Let $U = \{1, \ldots, N\}$ a finite population composed of N different units and let $s = \{1, 2, \ldots, n\}$ a random sample of size n selected using a specified sampling design $p(\cdot)$ with first and second-order inclusion probabilities $\pi_k > 0$ and $\pi_{kl} > 0$ $k, l \in U$ respectively and $d_k = \pi_k^{-1}$ denotes the sampling design-basic weight for unit $k \in U$. Let y_k be the study variable and $\mathbf{x}_k^{'} = (x_{1k}, \ldots, x_{Jk})$ be a vector of auxiliary variables at unit k. We assume that value $\mathbf{x}_k$ is available for all population units whereas the value y_k is available only for sample units. The distribution function $F_y(t)$ for the study variable y is defined as follow:

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k) \tag{4.1}$$

with

$$\Delta(t - y_k) = \begin{cases} 1 & \text{si } t \geq y_k \\ 0 & \text{si } t < y_k. \end{cases}$$

A design-based estimator of the distribution function $F_y(t)$ is the Horvitz–Thompson estimator, defined by

$$\widehat{F}_{YHT}(t) = \frac{1}{N} \sum_{k \in s} d_k \Delta(t - y_k). \tag{4.2}$$

The estimator $\widehat{F}_{YHT}(t)$ is unbiased, but it does not incorporate the auxiliary information provided by the auxiliary vector $\mathbf{x}$.

Several authors (Rueda et al., 2007a; Harms & Duchesne, 2006; Singh et al., 2008; Arcos et al., 2017) have incorporated the auxiliary information to obtain new estimators of $F_y(t)$ through the calibration method (Deville & Särndal, 1992). The proposalRueda et al. (2007a) applies the calibration procedure from a pseudo-variable

$$g_k = (\widehat{\beta})' \mathbf{x}_k \text{ for } k = 1, 2, ...N \tag{4.3}$$

$$\widehat{\beta} = \left(\sum_{k \in s} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \cdot \sum_{k \in s} d_k \mathbf{x}_k y_k \tag{4.4}$$

With the variable g, the basic weights d_k are replaced by new calibrated weights ω_k through the minimization of the chi-square distance measure

$$\Phi_s = \sum_{k \in s} \frac{(\omega_k - d_k)^2}{d_k q_k} \tag{4.5}$$

subject to the calibration constrains

$$\frac{1}{N} \sum_{k \in s} \omega_k \Delta(t_j - g_k) = F_g(t_j) \quad j = 1, 2, \ldots, P \tag{4.6}$$

where $F_g(t_j)$ denotes the finite distribution function of the pseudo-variable g_k evaluated at the points $t_j, \quad j = 1, 2, \ldots, P$. We assume, with no loss in generality, $t_1 < t_2 < \ldots t_P$. The values q_k are known positive constants unrelated to d_k.

Following Rueda et al. (2007a), we assume that the matrix T given by:

$$T = \sum_{k \in s} d_k q_k \Delta(\mathbf{t_g} - g_k) \Delta(\mathbf{t_g} - g_k)'$$

is nonsingular. With this calibration procedure, the calibration estimator obtained is:

$$\widehat{F}_{yc}(t) = \widehat{F}_{YHT}(t) + \left(F_g(\mathbf{t_g}) - \widehat{F}_{GHT}(\mathbf{t_g}) \right)' \cdot \widehat{D}(\mathbf{t_g}) \tag{4.7}$$

where

$$\widehat{D}(\mathbf{t_g}) = T^{-1} \cdot \sum_{k \in s} d_k q_k \Delta(\mathbf{t_g} - g_k) \Delta(t - y_k)$$

and $\widehat{F}_{GHT}(\mathbf{t_g})$ is the Horvitz-Thompson estimator of $F_g(\mathbf{t_g})$ evaluated at $\mathbf{t_g} = (t_1, \ldots, t_P)'$.

The calibration estimator $\widehat{F}_{yc}(t)$ has the following asymptotic variance (Rueda et al., 2007a):

$$AV(\widehat{F}_{yc}(t)) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k E_k)(d_l E_l) \tag{4.8}$$

where $E_k = \Delta(t - y_k) - \Delta(\mathbf{t_g} - g_k)\cdot D(\mathbf{t_g})$, with

$$D(\mathbf{t_g}) = \left(\sum_{k\in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)'\right)^{-1} \cdot \left(\sum_{k\in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(t - y_k)\right). \tag{4.9}$$

As a consequence, the behavior of the estimator $\widehat{F}_{yc}(t)$ and its precision depends on the selection of the vector $\mathbf{t_g}$.

Previous works Martínez et al. (2017, 2011, 2015) treated, under simple random sampling without replacement and $q_k = c$ for all $k \in U$, the optimal selection of the vector $\mathbf{t_g}$ in order to minimize the asymptotic variance (4.8). In fact, Martínez et al. (2017) established the optimal dimension of $\mathbf{t_g}$ and its optimal value, for a given value t, through the definition of the sets:

$$A_t = \{g_k : k \in U; y_k \leq t\} = \{a_1^t, a_2^t, \ldots, a_{M_t}^t\} \quad \text{with} \quad a_h^t < a_{h+1}^t \quad \text{for} \quad h = 1, \ldots, M_t - 1 \tag{4.10}$$

where M_t is the number of elements in the set A_t and

$$B_t = \{b_1^t, b_2^t, \ldots, b_{M_t}^t\} \tag{4.11}$$

with

$$b_1^t = \text{máx}_{l\in U_1}\{g_l\} \quad \text{where } U_1 = \{l \in U : g_l < a_1^t\}$$
$$b_h^t = \text{máx}_{l\in U_h}\{g_l\} \quad \text{where } U_h = \{l \in U : a_{h-1}^t < g_l < a_h^t\} \quad h = 2, 3, \ldots, M_t$$

and $b_h^t < b_{h+1}^t$ for $h = 1, \ldots, M_t - 1$.

Thus, Martínez et al. (2017) established that the auxiliary vector $\mathbf{t}_g$ has optimal dimension $P = 2M_t$ if b_h^t exists for $h = 1, \ldots, M_t$ and the optimal value of $\mathbf{t}_g$ is given by

$$\mathbf{t_{OPT}}(t) = (b_1^t, a_1^t, \ldots, b_{M_t}^t, a_{M_t}^t). \tag{4.12}$$

If there are some values $j_1^t, j_2^t, \ldots j_{p_t}^t \in \{1, \ldots, M_t\}$; such as $b_{j_h}^t$ does not exits for $h = 1, 2, \ldots p_t$ with $p_t \leq M_t$ and $j_h^t \neq j_q^t$ if $h \neq q$, the optimal dimension is given by $P = 2M_t - p_t$ and the optimal auxiliary

vector $\mathbf{t_{OP}}$ is:

$$\mathbf{t_{OP}}(t) = (b_1^t, a_1^t \ldots, b_{j_1-1}^t, a_{j_1-1}^t, a_{j_1}^t, b_{j_1+1}^t, \ldots, b_{j_h-1}^t, a_{j_h-1}^t, a_{j_h}^t, b_{j_h+1}^t, \ldots b_{M_t}^t, a_{M_t}^t). \tag{4.13}$$

In the next section, we will analyze if the minimum of the asymptotic variance can be reached with a vector of less dimension and we will establish conditions under which the dimension of the optimal vector $\mathbf{t_{OPT}}(t)$, can be reduced under simple random sampling without replacement.

3. Dimension reduction of the optimal auxiliary vector

In this section, we will analyze the conditions under which the dimension of the optimal vector $\mathbf{t_{OPT}}(t)$ can be reduced, that is, we will analyze the existence of a vector with a smaller dimension than $\mathbf{t_{OPT}}(t)$ that allows obtaining the minimum value of the asymptotic variance of the estimator $\widehat{F}_{yc}(t)$.

For the minimization of the asymptotic variance (4.8), we consider it as a function of a vector $\gamma = (\gamma_1, \ldots, \gamma_P)$ of dimension P:

$$AV(\widehat{F}_{yc}(t)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (d_k \Gamma_k)(d_l \Gamma_l) \tag{4.14}$$

with $\Gamma_k = \Delta(t - y_k) - \Delta(\gamma - g_k) \cdot D(\gamma)$, with $D(\gamma)$ given by (4.9).

Following Martínez et al. (2015), under simple random sampling without replacement and $q_k = c$ for all units in the population, the minimization of (4.14) is equivalent to the minimization of the function:

$$Q_t(\gamma) = Q_t(\gamma_1, \ldots, \gamma_P) = 2NF_y(t) \cdot K_t(\gamma_P) - \sum_{j=1}^{P} \frac{\left(K_t(\gamma_j) - K_t(\gamma_{j-1})\right)^2}{(F_g(\gamma_j) - F_g(\gamma_{j-1}))} - \left(K_t(\gamma_P)\right)^2 \tag{4.15}$$

with $K_t(\gamma_j) = \sum_{k \in U} \Delta(\gamma_j - g_k)\Delta(t - y_k)$, where we suppose that $F_g(\gamma_0) = 0$ and $K_t(\gamma_0) = 0$.

As mentioned above, Martínez et al. (2017) established the optimal dimension P and the minimum of (4.14) is reached at $\gamma = \mathbf{t_{OP}}(t)$. However, there are cases where the optimal dimension has a high value so the calibration procedure has a plenty of constraints which raises the computational cost for calculating the estimator. For example, if we consider $t = y_{max}$ where

$$y_{max} = \max_{k \in U} y_k$$

the optimal auxiliary vector $\mathbf{t_{OP}}(t) = (a_1, a_2, \ldots, a_M)$ can be reduced to the auxiliary vector $\gamma = (a_M)$ (see Appendix 1.1). Consequently, the optimal dimension can be reducted from M to 1.

In a similar way, we try to reduce the dimension of the auxiliary vector to reach the minimum of $Q_t(\gamma)$. For it, given a value t for which we want to estimate $F_y(t)$, we consider the sets A_M, A_t and B_t given by (26); (4.10) and (4.11) respectively and for each $a_i \in A_M$ we define:

$$r_i = \text{Frequency of the } a_i$$

For the value t, we have:

$$A_t = \{a_1^t, a_2^t, \ldots, a_{M_t}^t\} = \{a_{f_1^t}, a_{f_2^t}, \ldots, a_{f_{M_t}^t}\}$$

where

$$\{f_1^t, f_2^t, \ldots, f_{M_t}^t\} \subseteq \{1, 2, \ldots, M\} \text{ and } f_1^t < f_2^t < \cdots < f_{M_t}^t.$$

Similarly, we consider the following set:

$$C_t = \{g_k : k \in U; y_k > t\} = \{c_1^t, c_2^t, \ldots, c_{S_t}^t\} = \{a_{l_1^t}, a_{l_2^t}, \ldots, a_{l_{S_t}^t}\}$$

with

$$\{l_1^t, l_2^t, \ldots, l_{S_t}^t\} \subseteq \{1, 2, \ldots, M\} \text{ and } l_1^t < l_2^t < \cdots < l_{S_t}^t.$$

It is clear that $A_t \cup C_t = A_M$ and since for two different units k and j can be possible that $g_j = g_k = a_i$ and $y_k > t$ and $y_j < t$, not necessarily $A_t \cap C_t = \emptyset$. For the sets A_t and C_t we define:

$$p_i^t = \text{Frequency of the } a_i^t \text{ in } A_t$$
$$q_i^t = \text{Frequency of the } c_i^t \text{ in } C_t.$$

Next, we consider the following sets:

$$D_t = \{c_i \in C_t : q_i^t = r_i\} \tag{4.16}$$

$$Z_t = \{a_i^t \in A_t : q_i^t = 0\} = \{a_i^t \in A_t : a_i^t \notin C_t\} = A_t - C_t \tag{4.17}$$

$$F_t = \{a_i^t \in A_t : 0 < q_i^t < r_i\}. \tag{4.18}$$

It is easy to see that $D_t = A_M - A_t$ and consequently $A_t \cap D_t = \emptyset$. Futhermore, $B_t \subseteq D_t$; $A_t = Z_t \cup F_t$ and $Z_t \cap F_t = \emptyset$.

Firstly, if we suppose that $D_t = A_M$, we have $A_t = \emptyset$ and consequently $y_k > t$, $\forall k \in U$. In this case $F_y(t) = 0$ and we can calibrate with any auxiliary vector since

$$\widehat{F}_{yc}(t) = \frac{1}{N}\sum_{k\in s}\omega_k\Delta(t - y_k) = 0$$

regardless of the auxiliary vector, so we can calibrate with $\mathbf{t_{OP}}(t) = a_M$ with dimension 1.

Secondly, if we suppose that $D_t = \emptyset$, then $B_t = \emptyset$ and $A_t = A_M$. In this case, following Martínez et al. (2017) the optimal auxiliary vector $\mathbf{t_{OP}}(t) = (a_1, a_2, \ldots, a_M)$.

Since $A_t = Z_t \cap F_t = A_M$, if we suppose that $Z_t = A_t = A_M$ then $t > y_k$ $\forall k \in U$ and this case is like the case where $t = y_{max}$ and although the optimal auxiliary vector is $\mathbf{t_{OP}}(t) = (a_1, a_2, \ldots, a_M)$, we can reach the minimum value of $Q_\gamma(t)$ with the auxiliary vector $\gamma = (a_M)$.

On the other hand, if we consider that $Z_t = \emptyset$ and $F_t = A_M$ there is not reduction in the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ (see Appendix 1.2).

Next, if we suppose that $Z_t \neq A_t = A_M$ and $F_t \neq A_t = A_M$ then there is a set $I_{F_t} = \{j_1, j_2, \ldots, j_l\} \subseteq \{1, 2, \ldots M\}$ such that $a_{j_i} \in F_t$ and therefore $q^t_{j_i} \neq 0$ for $i = 1, 2, \ldots, l$.

Now, if we consider that $j_1 > 1$; $j_i - 1 > j_{(i-1)}$ for all $i = 2, \ldots l$ and $j_l < M$; then for $h = 1, \ldots j_1 - 1$; $q^t_h = 0$ and we have:

$$K_t(a_1) = \sum_{k\in U}\Delta(a_1 - g_k)\Delta(t - y_k) = NF_g(a_1)$$
$$\vdots$$
$$K_t(a_{(j_1-1)}) = \sum_{k\in U}\Delta(a_{j_1-1} - g_k)\Delta(t - y_k) = NF_g(a_{j_1-1}).$$

Similarly, for $j_i, \ldots j_{i+1} - 1$ with $i = 1, 2, \ldots l - 1$ we have:

$$K_t(a_{j_i}) = \sum_{k\in U}\Delta(a_{j_i} - g_k)\Delta(t - y_k) = NF_g(a_{j_i}) - \sum_{h=1}^{i} q^t_{j_h}$$
$$\vdots$$
$$K_t(a_{(j_{(i+1)}-1)}) = \sum_{k\in U}\Delta(a_{(j_{(i+1)}-1)} - g_k)\Delta(t - y_k) = NF_g(a_{(j_{(i+1)}-1)}) - \sum_{h=1}^{i} q^t_{j_h}$$

and finally, for $j_l, \ldots, M$

$$K_t(a_{j_l}) = \sum_{k\in U} \Delta(a_{j_l} - g_k)\Delta(t - y_k) = NF_g(a_{j_l}) - \sum_{h=1}^{l} q^t_{j_h}$$

$$\vdots$$

$$K_t(a_M) = \sum_{k\in U} \Delta(a_M - g_k)\Delta(t - y_k) = NF_g(a_M) - \sum_{h=1}^{l} q^t_{j_h} = NF_y(t).$$

The minimum of $Q_t(\gamma)$ reached at the optimum auxiliary vector $\mathbf{t_{OP}}(t)$ is given by:

$$Q_t(\mathbf{t_{OP}}(t)) = (NF_y(t))^2 - \sum_{j=1}^{M} \frac{\left(K_t(a_j - K_t(a_{j-1}))\right)^2}{F_g(a_j) - F_g(a_{j-1})} =$$

$$= (NF_y(t))^2 - N^2 \cdot \sum_{\substack{j=1 \\ j\notin\{j_1,\dots,j_l\}}}^{M} \frac{\left(F_g(a_j) - F_g(a_{j-1}))\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \sum_{j\in\{j_1,\dots,j_l\}} \frac{\left(NF_g(a_j) - NF_g(a_{j-1})) - q^t_j\right)^2}{F_g(a_j) - F_g(a_{j-1})}$$

$$= (NF_y(t))^2 - N^2 \cdot \sum_{\substack{j=1 \\ j\notin I_{F_t}}}^{M} \left(F_g(a_j) - F_g(a_{j-1}))\right) - N^2 \cdot \sum_{j\in I_{F_t}} \left(F_g(a_j) - F_g(a_{j-1})\right) + 2N \sum_{j\in I_{F_t}} q^t_j - \sum_{j\in I_{F_t}} \frac{\left(q^t_j)\right)^2}{F_g(a_j) - F_g(a_{j-1})} =$$

$$= (NF_y(t))^2 - N^2 \cdot \sum_{j=1}^{M} \left(F_g(a_j) - F_g(a_{j-1}))\right) + 2N \sum_{j\in I_{F_t}} q^t_j - \sum_{j\in I_{F_t}} \frac{\left(q^t_j\right)^2}{F_g(a_j) - F_g(a_{j-1})} =$$

$$= (NF_y(t))^2 - N^2 + 2N \sum_{j\in I_{F_t}} q^t_j - \sum_{j\in I_{F_t}} \frac{\left(q^t_j\right)^2}{F_g(a_j) - F_g(a_{j-1})}. \tag{4.19}$$

The same value can be reached with the auxiliary vector $\gamma = (a_{(j_1-1)}, a_{j_1}, \dots a_{(j_l-1)}, a_{j_l}, a_M)$. To see it, if we set a_{j0} such as $F_g(a_{j0}) = 0$ and we replace the vector γ in (4.15), we have:

$$Q_t(\gamma) = (N \cdot F_y(t))^2 - \sum_{h=1}^{l} \frac{\left(NF_g(a_{(j_h-1)}) - NF_g(a_{j_{(h-1)}})\right)^2}{F_g(a_{(j_h-1)}) - F_g(a_{j_{(h-1)}})} - \sum_{h=1}^{l} \frac{\left(NF_g(a_{j_h}) - NF_g(a_{(j_h-1)}) - q^t_{j_h})\right)^2}{F_g(a_{j_h}) - F_g(a_{(j_h-1)})}$$

$$-\frac{\left(NF_g(a_{(M)}) - NF_g(a_{j_l})\right)^2}{F_g(a_M) - F_g(a_{j_l})} = (N\cdot F_y(t))^2 - N^2 \sum_{h=1}^{l} F_g(a_{(j_h-1)}) - F_g(a_{j_{(h-1)}}) - N^2 \sum_{h=1}^{l} F_g(a_{j_h}) - F_g(a_{(j_h-1)})$$

$$+2N \sum_{h=1}^{l} q^t_{j_h} - \sum_{h=1}^{l} \frac{\left(q^t_{j_h}\right)^2}{F_g(a_{j_h}) - F_g(a_{(j_h-1)})} - N^2(F_g(a_M) - F_g(a_{j_l}))$$

$$= 2N \sum_{h=1}^{l} q^t_{j_h} - \sum_{h=1}^{l} \frac{\left(q^t_{j_h}\right)^2}{F_g(a_{j_h}) - F_g(a_{(j_h-1)})}$$

and the auxiliary vector γ, with less dimension than $\mathbf{t_{OP}}(t)$, attain the minimum of $Q_t(\gamma)$.

Previously, we suppose that $a_{j_1} > a_1$ and $a_{j_l} < a_M$. If $a_{j_1} = a_1$ then it is easy to see that the minimum can be obtain at $\gamma = (a_1, a_{(j_2-1)}, a_{j_2}, \ldots a_{(j_l-1)}, a_{j_l}, a_M)$ that has less dimension than in the case $a_{j_1} > a_1$. In a similar way, if $a_{j_l} = a_M$ the minimum can be attained at $\gamma = (a_{j_1}, a_{(j_1-1)}, a_{j_2}, \ldots a_{(j_l-1)}, a_M)$. Finally, we have assumed that $j_i - 1 > j_{(i-1)}$ for all $i = 2, \ldots l$. If there is a $h \in \{1, 2, \ldots l\}$ that $j_h - 1 = j_{(h-1)}$ it is easy to see that the minimum value of $Q_t(\gamma)$ is reached at $\gamma = (a_{(j_1-1)}, a_{j_1}, \ldots a_{j_{(h-1)}}, a_{j_h}, \ldots, a_{(j_l-1)}, a_{j_l}, a_M)$ with less dimension than in the case $j_i - 1 > j_{(i-1)}$ for all $i = 2, \ldots l$. Therefore, if $D_t = \emptyset$ we can reduce the optimal dimension when $F_t \neq A_t = A_M$.

Next, we consider the case where $D_t \neq \emptyset$ and $D_t \neq A_M$. Because $A_t = A_M - D_t$, we have $A_t \neq \emptyset$ and $A_t \neq A_M$. Therefore:

$$A_t = \{a_1^t, a_2^t, \ldots, a_{M_t}^t\} = \{a_{f_1^t}, a_{f_2^t}, \ldots, a_{f_{M_t}^t}\}$$

where $\{f_1^t, f_2^t, \ldots, f_{M_t}^t\} \subseteq \{1, 2, \ldots M\}$.
In this case, if we suppose that $B_t = \emptyset$ then $f_1^t = 1, \ldots, f_{M_t}^t = M_t$ and

$$A_t = \{a_1, a_2, \ldots a_{M_t}\} \quad ; \quad D_t = \{a_{M_t+1}, \ldots, a_M\}.$$

To see it, if we suppose that $f_1^t > 1$ then $a_{f_1^t} > a_{(f_1^t-1)} \geq a_1$ and consequently the set

$$U_1 = \{l \in U : g_l < a_1^t\} = \{l \in U : g_l < a_{f_1^t}\} \neq \emptyset$$

and $b_1^t = a_{(f_1^t-1)}$. Thus, $B_t \neq \emptyset$ (Contradiction). As a consequence, $a_{f_1^t} = a_1$.

If we suppose that for $i \in \{2, \ldots, M_t\}$ such as $f_{(i-1)}^t = i - 1$ and we suppose that $f_i^t > i$ then $a_{f_{(i-1)}^t} = a_{(i-1)}$ and $a_{f_i^t} > a_i > a_{(i-1)} = a_{f_{(i-1)}^t}$. The set U_i is given by:

$$U_i = \{l \in U : a_{(i-1)}^t < g_l < a_i^t\} = \{l \in U : a_{f_{(i-1)}} < g_l < a_{f_i}\} \neq \emptyset$$

and $b_i^t = a_{(f_i^t-1)}$. Thus, $B_t \neq \emptyset$ (Contradiction again). As a consequence, if $f_{(i-1)}^t = i - 1$ implies that $f_i^t = i$ for $i \in \{2, \ldots, M_t\}$ and we have:

$$A_t = \{a_1, a_2, \ldots a_{M_t}\}$$

If $M_t = M$ it is clear that $A_t = A_M$ and $D_t = \emptyset$ (Contradiction again). Therefore, $M_t < M$ and

$$D_t = A_M - A_t = \{a_{(M_t+1)}, a_{(M_t+2)}, \ldots a_M\}$$

The optimal auxiliary vector is given by $\mathbf{t_{OP}}(t) = (a_1, \ldots, a_{M_t})$ and in a similar way that in the previous cases, we can proof that if $F_t = A_t$, there is not a reduction in the optimal dimension. If $Z_t = A_t$, then we can attain the minimum of $Q_t(\gamma)$ at $\gamma = (a_{M_t})$. Finally, if $Z_t \neq A_t$ and $F_t \neq A_t$, and we suppose that

$$F_t = \{a_{j_1}, a_{j_2}, \ldots a_{jl}\}$$

we can reached the minimum value of $Q_\gamma(t)$ at $\gamma = (a_{(j_1-1)}, a_{j_1}, \ldots, a_{(j_l-1)}, a_{j_l}, a_{M_t})$.

Next, under the assumption $D_t \neq \emptyset$ and $D_t \neq A_M$, we consider $B_t \neq \emptyset$. If we assume that b_h^t exists for $h = 1, \ldots, M_t$, following the proposal Martínez et al. (2017), the optimal auxiliary vector $\mathbf{t}_{OPT}(t)$ is given by (4.12) and there is not a possible reduction in the dimension. On the other hand, if there are some values $j_1^t, j_2^t, \ldots j_{p_t}^t \in \{1, \ldots, M_t\}$; such as $b_{j_h}^t$ does not exits for $h = 1, 2, \ldots p_t$ with $p_t \leq M_t$ and $j_h^t \neq j_q^t$ if $h \neq q$, the optimal dimension is given by $P = 2M_t - p_t$ and the optimal auxiliary vector $\mathbf{t_{OP}}$ is given by (4.13). Analogously, there are $p_1, p_2, \ldots, p_{l_t} \in \{1, 2, \ldots, M_t\}$ such as $b_{f_{p_h}}^t$ exists and following Martínez et al. (2017) there is not a reduction between the points $b_{f_{p_h}}^t$ and $a_{f_{p_h}}^t$. Alternatively, the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ can be expressed as follows

$$\mathbf{t_{OP}}(t) = (a_{f_1^t}, \ldots, a_{f_{(p_1-1)}^t}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, a_{f_{(p_1+1)}^t}, \ldots, a_{f_{(p_h-1)}^t}, b_{f_{p_h}^t}, a_{f_{p_h}^t}, a_{f_{(p_h+1)}^t}, \ldots, a_{f_{M_t}^t}) \tag{4.20}$$

If we suppose that $p_1 = 1$ then the value $b_{f_1}^t$ exists and consequently:

$$U_1 = \{l \in U : g_l < a_1^t\} = \{l \in U : g_l < a_{f_1^t}\} \neq \emptyset$$

then, $a_{f_1^t} > a_1$ and $f_1^t > 1$. As a consequence $a_i \notin A_t$ with $i = 1, \ldots f_1^t - 1$ and $\{a_1, \ldots a_{(f_1^t-1)}\} \subseteq D_t$ which implies that $b_{f_1^t} = a_{(f_1^t-1)}$. In this case, we have:

$$K_t(b_{f_1^t}) = \sum_{k \in U} \Delta(b_{f_1^t} - g_k)\Delta(t - y_k) = 0$$

$$K_t(a_{f_1^t}) = \sum_{k \in U} \Delta(a_{f_1^t} - g_k)\Delta(t - y_k) = NF_g(a_{f_1^t}) - \sum_{h=1}^{f_1^t - 1} r_h - q_1^t = NF_g(a_{f_1^t}) - H_1^t$$

and following Martínez et al. (2017) there is no possibility to reduce the number of points here.

On the other hand, if we suppose that $p_1 > 1$, for $i \in \{1, 2, \ldots p_1 - 1\}$ the value $b_{f_i}^t$ does not exist and:

$$U_1 = \{l \in U : g_l < a_1^t\} = \{l \in U : g_l < a_{f_1^t}\} = \emptyset$$

$$U_i = \{l \in U : a^t_{(i-1)} < g_l < a^t_i\} = \{l \in U : a_{f^t_{(i-1)}} < g_l < a_{f^t_i}\} = \emptyset \text{ for } i = 2, \ldots, p_1 - 1$$

therefore $a_{f^t_1} = a_1$ and $a_{f^t_i} = a_{(f_{(i-1)}+1)}$ for $i = 2, \ldots, p_1 - 1$ and then:

$$a_{f^t_1} = a_1; a_{f^t_2} = a_2; \ldots; a_{f^t_{(p_1-1)}} = a_{(p_1-1)}.$$

Thus, if $p_1 > 1$ we have $\{a_1, \ldots, a_{(p_1-1)}\} \subseteq A_t$ and consequently the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ given by (4.20) can be expressed as follows:

$$\mathbf{t_{OP}}(t) = (a_1, \ldots, a_{(p_1-1)}, b_{f^t_{p_1}}, a_{f^t_{p_1}}, \mathbf{t}_R) =$$

where $\mathbf{t}_R$ denotes

$$\mathbf{t}_R = (a_{f^t_{(p_1+1)}}, \ldots, a_{f^t_{(p_h-1)}}, b_{f^t_{p_h}}, a_{f^t_{p_h}}, a_{f^t_{(p_h+1)}}, \ldots, a_{f^t_{M_t}}).$$

On the contrary, for p_1 we have

$$U_{p_1} = \{l \in U : a^t_{(p_1-1)} < g_l < a^t_{p_1}\} = \{l \in U : a_{f^t_{(p_1-1)}} < g_l < a_{f^t_{p_1}}\} =$$

$$= \{l \in U : a_{(p_1-1)} < g_l < a_{f^t_{p_1}}\} \neq \emptyset$$

and then $a_{f^t_{p_1}} > a_{p_1}$ and $f^t_{p_1} > p_1$ which implies that

$$\{a_{p_1}, a_{(p_1+1)}, \ldots a_{(f^t_{p_1}-1)}\} \subseteq D_t.$$

We consider the following sets:

$$A_{p_1} = \{a_1, \ldots, a_{(p_1-1)}\} \tag{4.21}$$

$$Z_{p_1} = \{a_i \in A_{p_1} : q^t_i = 0\} \tag{4.22}$$

$$F_{p_1} = \{a_i \in A_{p_1} : 0 < q^t_i < r_i\} \tag{4.23}$$

Similarly to the previous cases, if we suppose that $Z_{p_1} = A_{p_1}$ and $F_{p_1} = \emptyset$, it is easy to see that we can delete a_i for $i = 1, 2, \ldots p_1 - 2$ from the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ and we can calibrate only with the value $a_{(p_1-1)}$. To see it, it is clear that:

$$K_t(a_1) = \sum_{k\in U} \Delta(a_1 - g_k)\Delta(t - y_k) = NF_g(a_1)$$
$$\vdots$$
$$K_t(a_{(p_1-1)}) = \sum_{k\in U} \Delta(a_{j_1-1} - g_k)\Delta(t - y_k) = NF_g(a_{(p_1-1)}).$$

Because $\{a_{p_1}, a_{(p_1+1)}, \dots a_{(f_{p_1}-1)}\} \subseteq D_t$ it is easy to see that:

$$K_t(b_{f^t_{p_1}}) = \sum_{k\in U} \Delta(b_{f^t_{p_1}} - g_k)\Delta(t - y_k) = K_t(a_{(p_1-1)}) = NF_g(a_{(p_1-1)})$$
$$K_t(a_{f^t_{p_1}}) = \sum_{k\in U} \Delta(a_{f^t_{p_1}} - g_k)\Delta(t - y_k) = NF_g(a_{f^t_{p_1}}) - \sum_{h=p_1}^{f_{p_1}-1} r_h - q^t_{f_{p_1}} = NF_g(a_{f^t_{p_1}}) - H^t_{p_1}.$$

The minimum value of $Q_t(\gamma)$ at $\mathbf{t_{OP}}(t)$ can be expressed as follows:

$$Q_t(\mathbf{t_{OP}}(t)) = Q_t(a_1, \dots, a_{(p_1-1)}, b_{f^t_{p_1}}, a_{f^t_{p_1}}, \mathbf{t}_R) =$$

$$= Q_t(\mathbf{t}_R) - \sum_{j=1}^{p_1-1} \frac{\left(K_t(a_j) - K_t(a_{j-1})\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \frac{\left(K_t(b_{f^t_{p_1}}) - K_t(a_{(p_1-1)})\right)^2}{F_g(b_{f^t_{p_1}}) - F_g(a_{(p_1-1)})} - \frac{\left(K_t(a_{f^t_{p_1}}) - K_t(b_{f^t_{p_1}})\right)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})} =$$

$$= Q_t(\mathbf{t}_R) - N^2 \sum_{j=1}^{p_1-1} \left(F_g(a_j) - F_g(a_{j-1})\right) - \frac{\left(NF_g(a_{f^t_{p_1}}) - NF_g(a_{(p_1-1)}) - H^t_{p_1}\right)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})} =$$

$$= Q_t(\mathbf{t}_R) - N^2 F_g(a_{(p_1-1)}) - \frac{\left(NF_g(a_{f^t_{p_1}}) - NF_g(a_{(p_1-1)}) - H^t_{p_1}\right)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})}.$$

If we calibrate with the auxiliary vector $\gamma = (a_{(p_1-1)}, b_{f^t_{p_1}}, a_{f^t_{p_1}}, \mathbf{t}_R)$, we obtain the same value:

$$Q_t(\gamma) = Q_t(a_{(p_1-1)}, b_{f^t_{p_1}}, a_{f^t_{p_1}}, \mathbf{t}_R) =$$

$$= Q_t(\mathbf{t}_R) - \frac{\left(K_t(a_{(p_1-1)}))\right)^2}{F_g(a_{(p_1-1)})} - \frac{\left(K_t(b_{f^t_{p_1}}) - K_t(a_{(p_1-1)})\right)^2}{F_g(b_{f^t_{p_1}}) - F_g(a_{(p_1-1)})} - \frac{\left(K_t(a_{f^t_{p_1}}) - K_t(b_{f^t_{p_1}})\right)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})} =$$

$$= Q_t(\mathbf{t}_R) - N^2 F_g(a_{(p_1-1)}) - \frac{\left(NF_g(a_{f^t_{p_1}}) - NF_g(a_{(p_1-1)}) - H^t_{p_1}\right)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})}.$$

Thus, if $Z_{p_1} = A_{p_1}$ we can reduce the dimension of the auxiliary vector to attain the minimum value of $Q_t(\gamma)$.

Now, if we suppose that $F_{p_1} = A_{p_1}$ and $Z_{p_1} = \emptyset$, as in the previous cases, we cannot reduce the dimension of the subvector $(a_1, \dots, a_{(p_1-1)})$ in the auxiliary vector $(\mathbf{t_{OP}}(t))$.

Next, if we suppose that $Z_{p_1} \neq A_{p_1}$ and $F_{p_1} \neq A_{p_1}$ then there is a set $I_{F_{p_1}} = \{j^1_1, j^2_2, \dots, j^1_{l_1}\} \subseteq$

$\{1, 2, \ldots p_1 - 1\}$ such that $a_{j_h^1} \in F_{p_1}$ and therefore $q^t_{j_h^1} \neq 0$ for $h = 1, 2, \ldots, l_1$.

Now, if we consider that $j_1^1 > 1$; $j_h^1 - 1 > j^1_{(h-1)}$ for all $h = 2, \ldots l_1$ and $j^1_{l_1} < p_1 - 1$; then for $v = 1, \ldots j_1^1 - 1$; $q^t_v = 0$ and we have:

$$K_t(a_1) = \sum_{k\in U} \Delta(a_1 - g_k)\Delta(t - y_k) = NF_g(a_1)$$

$$\vdots$$

$$K_t(a_{(j_1^1-1)}) = \sum_{k\in U} \Delta(a_{j_1^1-1} - g_k)\Delta(t - y_k) = NF_g(a_{j_1^1-1})$$

Similarly, for $j_h^1, \ldots j^1_{h+1} - 1$ with $h = 1, 2, \ldots l_1 - 1$; we have:

$$K_t(a_{j_h^1}) = \sum_{k\in U} \Delta(a_{j_h^1} - g_k)\Delta(t - y_k) = NF_g(a_{j_h^1}) - \sum_{v=1}^{h} q^t_{j_v^1}$$

$$\vdots$$

$$K_t(a_{(j^1_{(h+1)}-1)}) = \sum_{k\in U} \Delta(a_{(j^1_{(h+1)}-1)} - g_k)\Delta(t - y_k) = NF_g(a_{(j^1_{(h+1)}-1)}) - \sum_{v=1}^{h} q^t_{j_v^1}$$

and finally, for $j^1_{l_1}, \ldots, p_1 - 1$

$$K_t(a_{j^1_{l_1}}) = \sum_{k\in U} \Delta(a_{j^1_{l_1}} - g_k)\Delta(t - y_k) = NF_g(a_{j^1_{l_1}}) - \sum_{v=1}^{l_1^1} q^t_{j_v^1} = NF_g(a_{j^1_{l_1}}) - L_1^t$$

$$\vdots$$

$$K_t(a_{(p_1-1)}) = \sum_{k\in U} \Delta(a_{(p_1-1)} - g_k)\Delta(t - y_k) = NF_g(a_{(p_1-1)}) - \sum_{v=1}^{l_1^1} q^t_{j_v^1} = NF_g(a_{(p_1-1)}) - L_1^t.$$

Again, because $\{a_{p_1}, a_{(p_1+1)}, \ldots a_{(f_{p_1}-1)}\} \subseteq D_t$ we have:

$$K_t(b_{f^t_{p_1}}) = \sum_{k\in U} \Delta(b_{f^t_{p_1}} - g_k)\Delta(t - y_k) = K_t(a_{(p_1-1)}) = NF_g(a_{(p_1-1)}) - L_1^t$$

$$K_t(a_{f^t_{p_1}}) = \sum_{k\in U} \Delta(a_{f^t_{p_1}} - g_k)\Delta(t - y_k) = NF_g(a_{f^t_{p_1}}) - L_1^t - \sum_{h=p_1}^{f_{p_1}-1} r_h - q^t_{f_{p_1}} = NF_g(a_{f^t_{p_1}}) - L_1^t - H^t_{p_1}$$

The minimum of $Q_t(\gamma)$ at $(\mathbf{t_{OP}}(t))$ is given by:

$$Q_t(\mathbf{t_{OP}}(t)) = Q_t(a_1, \ldots, a_{(p_1-1)}, b_{f^t_{p_1}}, a_{f^t_{p_1}}, \mathbf{t}_R) =$$

$$= Q_t(\mathbf{t}_R) - \sum_{j=1}^{p_1-1} \frac{\big(K_t(a_j) - K_t(a_{j-1})\big)^2}{F_g(a_j) - F_g(a_{j-1})} - \frac{\big(K_t(b_{f^t_{p_1}}) - K_t(a_{(p_1-1)})\big)^2}{F_g(b_{f^t_{p_1}}) - F_g(a_{(p_1-1)})} - \frac{\big(K_t(a_{f^t_{p_1}}) - K_t(b_{f^t_{p_1}})\big)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})} =$$

$$= Q_t(\mathbf{t}_R) - N^2 \sum_{j=1}^{p_1-1} \big(F_g(a_j) - F_g(a_{j-1})\big) + 2N \cdot L_1^t - \sum_{v=1}^{l_1^1} \frac{\big(q^t_{j_v^1}\big)^2}{F_g(a_{j_v^1}) - F_g(a_{(j_1^v-1)})} - \frac{\big(NF_g(a_{f^t_{p_1}}) - NF_g(a_{(p_1-1)}) - H^t_{p_1}\big)^2}{F_g(a_{f^t_{p_1}}) - F_g(b_{f^t_{p_1}})} =$$

$$= Q_t(\mathbf{t}_R) - N^2 F_g(a_{(p_1-1)}) + 2N \cdot L_1^t - \sum_{v=1}^{l_1^1} \frac{\left(q_{j_v^1}^t\right)^2}{F_g(a_{j_v^1}) - F_g(a_{(j_v^1-1)})} - \frac{\left(NF_g(a_{f_{p_1}^t}) - NF_g(a_{(p_1-1)}) - H_{p_1}^t\right)^2}{F_g(a_{f_{p_1}^t}) - F_g(b_{f_{p_1}^t})}.$$

If we calibrate with the following auxiliary vector:

$$\gamma = \left(a_{(j_1^1-1)}, a_{j_1^1}, \ldots, a_{(j_{l_1}^1-1)}, a_{j_{l_1}^1}, a_{(p_1-1)}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R\right)$$

in a similar way to the previous cases, it is easy to see that:

$$Q_t\left(a_{(j_1^1-1)}, a_{j_1^1}, \ldots, a_{(j_{l_1}^1-1)}, a_{j_{l_1}^1}, a_{(p_1-1)}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R\right) =$$

$$= Q_t(\mathbf{t}_R) - N^2 F_g(a_{(p_1-1)}) + 2N \cdot L_1^t - \sum_{v=1}^{l_1^1} \frac{\left(q_{j_v^1}^t\right)^2}{F_g(a_{j_v^1}) - F_g(a_{(j_v^1-1)})} - \frac{\left(NF_g(a_{f_{p_1}^t}) - NF_g(a_{(p_1-1)}) - H_{p_1}^t\right)^2}{F_g(a_{f_{p_1}^t}) - F_g(b_{f_{p_1}^t})}.$$

Then, if $Z_{p_1} \neq A_{p_1}$ and $F_{p_1} \neq A_{p_1}$ again we can reduce the dimension of the optimal vector $\mathbf{t_{OP}}(t)$.

Finally, we have assumed that $j_1^1 > 1$; $j_h^1 - 1 > j_{(h-1)}^1$ for all $h = 2, \ldots l_1$ and $j_{l_1}^1 < p_1 - 1$. If there is a $h \in \{1, 2, \ldots l\}$ that $j_h^1 - 1 = j_{(h-1)}^1$ it is easy to see that the minimum value of $Q_t(\gamma)$ is reached at

$$\gamma = \left(a_{(j_1^1-1)}, a_{j_1^1}, \ldots a_{j_{(h-1)}^1}, a_{j_h^1}, \ldots, a_{(j_{l_1}^1-1)}, a_{j_{l_1}^1}, a_{(p_1-1)}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R\right)$$

with less dimension than in the case $j_h^1 - 1 > j_{(h-1)}^1$ for all $h = 2, \ldots l_1$. Similarly, if we suppose that $j_1^1 = 1$ then the minimum is obtained with

$$\gamma = \left(a_{j_1^1}, a_{(j_2^1-1)}, a_{j_2^1}, \ldots a_{j_{(h-1)}^1}, a_{j_h^1}, \ldots, a_{(j_{l_1}^1-1)}, a_{j_{l_1}^1}, a_{(p_1-1)}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R\right)$$

again, with less dimension than in the case $j_1^1 > 1$. In a similar way, if $j_{l_1}^1 = p_1 - 1$, the minimum of $Q_t(\gamma)$ is reached at

$$\gamma = \left(a_{(j_1^1-1)}, a_{j_1^1}, \ldots a_{j_{(h-1)}^1}, a_{j_h^1}, \ldots, a_{(j_{l_1}^1-1)}, a_{j_{l_1}^1}, b_{f_{p_1}^t}, a_{f_{p_1}^t}, \mathbf{t}_R\right)$$

again, with less dimension than in the case $j_{l_1}^1 < p_1 - 1$.

Now, if we consider p_i with $i = 2, \ldots, l_t$, we can extend the analysis for the dimension reduction of the optimal auxiliary vector in a similar way to the case p_1.(see Appendix 1.3).

4. The new optimal estimator with the new optimal vector

In the previous section we have theoretically demonstrated that the optimal vector $\mathbf{t_{OPT}}(t)$ proposed in Martínez et al. (2017) can be reduced in its dimension and we can minimize the variance given by (4.8) with a new optimal vector $\mathbf{t_{NEWOPT}}(t)$ of lower dimension. As with the original optimal vector $\mathbf{t_{OPT}}(t)$, the new vector $\mathbf{t_{NEWOPT}}(t)$ depends on unknown population values and therefore needs to be estimated. For it, in the same way as in Martínez et al. (2017), from the sample versions of the sets A_t, B_t, C_t, D_t, Z_t and F_t and the sample version of the function $Q_t(\gamma)$, an estimate $\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t)$ of the vector $\mathbf{t_{NEWOPT}}(t)$ can be obtained and we can define a new calibrated estimator $\widehat{F}_{CALNEWOPT}(t)$ for the distribution function $F_y(t)$, given by:

$$\widehat{F}_{CALNEWOPT}(t) = \widehat{F}_{YHT}(t) + \left(F_g(\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t)) - \widehat{F}_{GHT}(\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t))\right)' \cdot \widehat{D}\left(\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t)\right) \tag{4.24}$$

where

$$\widehat{D}\left(\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t)\right) = T^{-1} \cdot \sum_{k \in s} d_k q_k \Delta(\widehat{\mathbf{t}}_{\mathbf{NEWOPT}}(t) - g_k)\Delta(t - y_k)$$

5. Simulation study

In this section, a simulation study was conducted to compare the performance of the proposed optimal estimator with other alternative estimators for the distribution function $F(t)$. The simulation study was programmed in R software [version 4.1.0] and it was necessary to develop a new code to calculate the estimators included in the simulation study. The precision of the proposed new optimal calibration estimator $\widehat{F}_{CALNEWOPT}(t)$ was compared with the following estimators, the Horvitz Thompson estimator, $\widehat{F}_{HT}$, the difference estimator (Rao et al., 1990), $\widehat{F}_D(t)$, the ratio estimator (Rao et al., 1990) $\widehat{F}_R(t)$, the chambers1986estimating-Dunstan estimator (Chambers & Dunstan, 1986) $\widehat{F}_{CD}(t)$, the Rao-Kovar-Mantel estimator (Rao et al., 1990) $\widehat{F}_{RKM}(t)$, the calibration estimator (Rueda et al., 2007a) with $t_1 = Q_g(0{,}5)$, the population median, as point for calibration, $\widehat{F}_{CAL}(t)$, the calibration estimator (Rueda et al., 2007a) with three points $t_1 = Q_g(0{,}25)$, $t_2 = Q_g(0{,}5)$ and $t_3 = Q_g(0{,}75)$, the population quartiles, as points for calibration, $\widehat{F}_{CAL,3}(t)$, the calibration estimator with one optimal point (Martínez et al., 2010), $\widehat{F}_{CALMAX}(t)$ and finally the previous optimal calibration estimator (Martínez et al., 2017) $\widehat{F}_{CALOPT}(t)$.

Both real populations and simulated populations were considered for the simulation study. Specifically, we considered a real population included in The R Datasets Package called DNase that provides data collected from an ELISA assay for recombinant DNase protein in rat serum with population size $N = 176$. In addition, two simulated population called Simh and Simser were considered. The first one, Simh, is a population of

size $N = 5000$ generated from the following superpopulation model:

$$y_k = 8 - 7{,}82/x_k + \epsilon_k$$

where x is a sample from a discrete uniform distribution in $\{1, 2, \ldots, 100\}$ and ϵ_k's are i.i.d. random variables from $N(0, 0{,}5/x_k)$.

The simulated population Simser is a population of size $N = 5882$ generated from the following superpopulation model:

$$y_k = x_k^2 + \epsilon_k$$

where x is a sample from a discrete uniform distribution in $\{-100, -99, \ldots, 100\}$ and ϵ_k's are i.i.d. random variables from $N(0, 10)$.

For each population included in the simulation study, we drawn by simple random sampling without replacement 1000 samples of several sizes. For each sample, we estimated the distribution function $F(t)$ through all the estimators considered in the study at 11 different values of t, namely the quantiles $Q_y(\alpha)$ for α=0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8 and 0.9.

To measure the performance of each estimator included in the study, we considered the average relative bias (AVRB) and the average relative efficiency (AVRE), defined as follow:

$$\text{AVRB}(t) = \frac{1}{11}\sum_{q=1}^{11} |\text{RB}(t_q)|, \quad \text{AVRE}(t) = \frac{1}{11}\sum_{q=1}^{11} \text{RE}(t_q)$$

where RB and RE are defined as

$$\text{RB}(t) = \frac{1}{B}\sum_{b=1}^{B} \frac{\widehat{F}(t)_b - F_y(t)}{F_y(t)} \quad \text{and} \quad \text{RE}(t) = \frac{MSE[\widehat{F}(t)]}{MSE[\widehat{F}_{HT}(t)]}, \tag{4.25}$$

where b indexes the bth simulation run, $\widehat{F}(t)$ is an estimator for the distribution function, $MSE[\widehat{F}(t)] = B^{-1}\sum_{b=1}^{B}[\widehat{F}(t)_b - F_y(t)]^2$ is the empirical mean square error for $\widehat{F}(t)$ and $MSE[\widehat{F}_{HT}(t)]$ is similarly defined for the Horvitz-Thompson estimator.

Given that the new estimator proposal $\widehat{F}_{CALNEWOPT}(t)$ and the estimator $\widehat{F}_{CALOPT}(t)$ are based on the minimization of (4.15), it is possible that their behavior in terms of efficiency is similar and therefore it is necessary to analyze their behavior in greater detail. A reduced dimensionality in the auxiliary information set may reduce numerical issues in optimization procedures and also avoid the presence of unstable calibration

weights (both negative weights and huge weights) . Therefore, for each of the eleven estimation points t_q, we compared the dimension of the optimum auxiliary vector used in each estimators $\widehat{F}_{CALOPT}(t)$ and $\widehat{F}_{CALNEWOPT}(t)$. For it, we considered the mean dimension and the variance of the dimension:

$$\text{MD}(\widehat{F}(t_q)) = \frac{1}{B}\sum_{b=1}^{B} \text{DIM}(\mathbf{t_{qopt}}), \quad \text{VD}(\widehat{F}(t_q)) = \frac{1}{B}\sum_{b=1}^{B} \left(\text{DIM}(\mathbf{t_{qopt}}) - \text{MD}t_q\right)^2$$

where $\widehat{F}$ can be $\widehat{F}_{CALOPT}(t)$ or $\widehat{F}_{CALNEWOPT}(t)$, $\mathbf{t_{qopt}}$ denote the optimum auxiliary vector used with the point t_q and $\text{DIM}(\mathbf{t_{qopt}})$ denote the dimension of $(\mathbf{t_{qopt}})$.

Additionally, because a limited number of variables may reduce the execution time to resolve the calibration procedure, we compared for each estimation point the execution time in calculating the estimators using the following measure:

$$\text{RT}(t) = \frac{\text{TIME}(\widehat{F}_{CALNEWOPT}(t))}{\text{TIME}(\widehat{F}_{CALOPT}(t))}$$

where $\text{TIME}(\widehat{F}_{CALOPT}(t))$ and $\text{TIME}(\widehat{F}_{CALNEWOPT}(t))$ denote the running time for calculating $\widehat{F}_{CALOPT}(t)$ and $\widehat{F}_{CALNEWOPT}(t)$ respectively.

For the population DNase, Table A4.1 gives the values of AVRB and AVRE whereas Table A4.2 gives the values of MD; VD and RT. With respect to the results obtained for the bias and efficiency analysis, the estimators $\widehat{F}_{CALOPT}(t)$ and $\widehat{F}_{CALNEWOPT}(t)$ show the same behavior. Thus, both estimators present an adequate bias value, although for all sample sizes there are estimators that present a lower bias. In relation to efficiency, both estimators are clearly the most efficient. From results in Table A4.2, as expected, the estimator $\widehat{F}_{CALNEWOPT}(t)$ always presents a smaller dimension of the auxiliary vector used, but this reduction is quite modest for all sample sizes and therefore the reduction obtained in the execution time is also quite modest. This may be because the set F_t has a cardinal similar to the set A_t or the set B_t also has a cardinal similar to the set A_t. Consequently, the new proposal $\widehat{F}_{CALNEWOPT}(t)$ only achieves small reductions in the dimension of the auxiliary information used with respect to $\widehat{F}_{CALOPT}(t)$, that produces slight improvements in execution time and it is not considerable enough to improve the asymptotic behavior (Chauvet & Goga, 2022), although it does not deteriorate it either.

Tables A4.3 and A4.4 provide the results obtained for the Simh population. From the results of Table A4.3 (AVRB and AVRE) we can again observe that $\widehat{F}_{CALOPT}(t)$ and $\widehat{F}_{CALNEWOPT}(t)$ have the same behavior and they are the most efficient estimators for all sample sizes. Also, they also present a less bias for most of sample sizes. Additionally, we can highlight the bias and efficiency problems of $\widehat{F}_{CD}(t)$ because this estimator is biased when the relationship between y and x is not linear. As in the previous case, the dimension reduction

Tabla A4.1: Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared. Population: DNase.

	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 30$		$n = 32$		$n = 35$		$n = 37$	
$\widehat{F}_{HT}$	0.0064	1	0.0063	1	0.0037	1	0.0030	1
$\widehat{F}_{D}$	0.0136	0.4898	0.0219	0.4867	0.0173	0.4642	0.0153	0.4703
$\widehat{F}_{R}$	0.0083	0.4412	0.0040	0.4343	0.0061	0.4247	0.0024	0.4335
$\widehat{F}_{CD}$	0.1869	0.8930	0.1878	0.9150	0.1814	0.9441	0.1783	0.9749
$\widehat{F}_{RKM}$	0.0032	0.4128	0.0103	0.4124	0.0063	0.4023	0.0055	0.4068
$\widehat{F}_{CAL}$	0.0066	0.8575	0.0090	0.8377	0.0029	0.8881	0.0012	0.8436
$\widehat{F}_{CAL3}$	0.0046	0.3401	0.0035	0.3468	0.0053	0.3322	0.0015	0.3261
$\widehat{F}_{CALMAX}$	0.0057	0.2102	0.0079	0.2247	0.0052	0.1981	0.0050	0.1991
$\widehat{F}_{CALOPT}$	0.0047	0.1895	0.0059	0.1928	0.0030	0.1676	0.0024	0.1629
$\widehat{F}_{CALNEWOPT}$	0.0047	0.1895	0.0059	0.1928	0.0030	0.1676	0.0024	0.1629
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 40$		$n = 42$		$n = 45$		$n = 47$	
$\widehat{F}_{HT}$	0.0056	1	0.0032	1	0.0044	1	0.0052	1
$\widehat{F}_{D}$	0.0160	0.4620	0.0102	0.4670	0.0089	0.4675	0.0108	0.4481
$\widehat{F}_{R}$	0.0018	0.4241	0.0033	0.4380	0.0045	0.4389	0.0041	0.4160
$\widehat{F}_{CD}$	0.1759	1.0417	0.1703	1.0957	0.1625	1.1100	0.1598	1.1054
$\widehat{F}_{RKM}$	0.0076	0.4012	0.0023	0.4070	0.0011	0.4136	0.0017	0.4014
$\widehat{F}_{CAL}$	0.0049	0.8386	0.0032	0.8394	0.0034	0.9319	0.0034	0.8700
$\widehat{F}_{CAL3}$	0.0008	0.3397	0.0024	0.3466	0.0033	0.3442	0.0022	0.3273
$\widehat{F}_{CALMAX}$	0.0043	0.1833	0.0028	0.1938	0.0031	0.1962	0.0036	0.1972
$\widehat{F}_{CALOPT}$	0.0022	0.1530	0.0013	0.1582	0.0011	0.1557	0.0019	0.1516
$\widehat{F}_{CALNEWOPT}$	0.0022	0.1530	0.0013	0.1582	0.0011	0.1557	0.0019	0.1516

Tabla A4.2: Average dimension(MD), variance dimension (VD) and comparison of execution time (RT) of the estimators $\widehat{F}_{CALOPT}$ and $\widehat{F}_{CALNEWOPT}$. Population: DNase.

	$n = 30$					$n = 32$				
	MD		VD		RT	MD		VD		RT
	OPT	*NEWOPT*	*OPT*	*NEWOPT*		*OPT*	*NEWOPT*	*OPT*	*NEWOPT*	
t_1	1	1	0	0	1	1	1	0	0	1
t_2	1.929	1.698	0.257	0.459	0.875	1.920	1.742	0.271	0.438	0.556
t_3	1.976	1	0.153	0	0.600	1.978	1	0.147	0	0.217
t_4	2.781	1.738	0.419	0.440	0.846	2.807	1.761	0.405	0.427	0.467
t_5	3.658	1.668	0.499	0.471	0.813	3.665	1.670	0.491	0.470	0.895
t_6	3.951	1	0.225	0	0.238	3.960	1	0.196	0	0.412
t_7	4.915	1.515	0.286	0.500	0.375	4.939	1.530	0.244	0.499	0.849
t_8	5.883	1.739	0.343	0.439	0.579	5.897	1.763	0.307	0.425	0.261
t_9	5.921	1	0.288	0	0.556	5.938	1	0.241	0	0.600
t_{10}	6.723	1.733	0.476	0.443	0.571	6.765	1.756	0.452	0.430	0.909
t_{11}	7.502	1.573	0.580	0.495	0.536	7.578	1.632	0.541	0.483	0.950
	$n = 35$					$n = 37$				
	MD		VD		RT	MD		VD		RT
	OPT	*NEWOPT*	*OPT*	*NEWOPT*		*OPT*	*NEWOPT*	*OPT*	*NEWOPT*	
t_1	1	1	0	0	1	1	1	0	0	0.187
t_2	1.951	1.816	0.216	0.388	0.972	1.960	1.815	0.196	0.388	0.974
t_3	1.989	1	0.104	0	0.379	1.990	1	0.100	0	0.250
t_4	2.853	1.803	0.357	0.398	0.762	2.876	1.847	0.330	0.360	0.941
t_5	3.714	1.714	0.465	0.452	0.650	3.749	1.742	0.438	0.438	0.615
t_6	3.977	1	0.150	0	0.375	3.986	1	0.118	0	0.833
t_7	4.962	1.608	0.196	0.488	0.524	4.971	1.572	0.168	0.495	0.773
t_8	5.936	1.817	0.249	0.387	0.515	5.948	1.803	0.231	0.398	0.599
t_9	5.969	1	0.173	0	0.125	5.975	1	0.162	0	0.471
t_{10}	6.823	1.810	0.382	0.392	0.769	6.865	1.864	0.353	0.343	0.905
t_{11}	7.650	1.671	0.494	0.470	0.667	7.654	1.677	0.497	0.468	0.714
	$n = 40$					$n = 42$				
	MD		VD		RT	MD		VD		RT
	OPT	*NEWOPT*	*OPT*	*NEWOPT*		*OPT*	*NEWOPT*	*OPT*	*NEWOPT*	
t_1	1	1	0	0	1	1	0	0	1	
t_2	1.983	1.862	0.129	0.345	0.760	1.982	1.889	0.133	0.314	0.982
t_3	1.998	1	0.0447	0	0.818	1.996	1	0.063	0	0.353
t_4	2.898	1.871	0.303	0.335	0.706	2.920	1.895	0.271	0.307	0.753
t_5	3.793	1.789	0.405	0.408	0.278	3.782	1.781	0.413	0.414	0.647
t_6	3.996	1	0.063	0	0.346	3.993	1	0.083	0	0.500
t_7	4.989	1.613	0.104	0.487	0.615	4.989	1.665	0.104	0.472	0.698
t_8	5.974	1.869	0.159	0.338	0.587	5.979	1.880	0.143	0.325	0.440
t_9	5.991	1	0.094	0	0	5.992	1	0.089	0.403	0.214
t_{10}	6.909	1.882	0.291	0.323	0.814	6.893	1.875	0.309	0.331	0.692
t_{11}	7.731	1.735	0.446	0.441	0.773	7.750	1.756	0.433	0.430	0.400
	$n = 45$					$n = 47$				
	MD		VD		RT	MD		VD		RT
	OPT	*NEWOPT*	*OPT*	*NEWOPT*		*OPT*	*NEWOPT*	*OPT*	*NEWOPT*	
t_1	1	1	0	0	1	1	1	0	0	1
t_2	1.989	1.905	0.104	0.293	0.991	1.993	1.920	0.083	0.271	0.882
t_3	1.998	1	0.045	0	0.467	1.999	1	0.032	0	0.600
t_4	2.929	1.914	0.257	0.281	0.793	2.937	1.929	0.243	0.257	0.733
t_5	3.834	1.829	0.372	0.377	0.882	3.863	1.863	0.347	0.344	0.805
t_6	3.994	1	0.077	0	0.083	3.997	1	0.055	0	0.556
t_7	4.990	1.698	0.100	0.459	0.786	4.995	1.705	0.071	0.456	0.529
t_8	5.981	1.908	0.137	0.289	0.450	5.982	1.920	0.133	0.271	0.857
t_9	5.993	1	0.083	0	0.136	5.995	1	0.071	0	0.368
t_{10}	6.928	1.919	0.262	0.273	0.875	6.931	1.918	0.254	0.275	0.615
t_{11}	7.790	1.794	0.415	0.405	0.782	7.782	1.784	0.416	0.412	0.643

analysis (Table A4.4) is essential to find out if $\widehat{F}_{CALNEWOPT}(t)$ is a better alternative than $\widehat{F}_{CALOPT}(t)$. In this case, from the results of Table A4.4, we can verify that again there is a slight reduction in the optimal vector used in $\widehat{F}_{CALNEWOPT}(t)$ for the smalls and medium quantiles. On the contrary, there is a moderate reduction for the higher quantiles where the dimension of the optimal vector used in $\widehat{F}_{CALOPT}(t)$ has a value between 10 and 20 while the dimension for $\widehat{F}_{CALNEWOPT}(t)$ remains between 2 and 5 for all sample sizes. Due to this reduction, the new estimator $\widehat{F}_{CALNEWOPT}(t)$ provides a considerable benefit in execution time, especially in the higher quantiles but as in the previous case, this moderate reduction in the dimension of the auxiliary information used in the calibration procedure does not allow an improvement in the asymptotic efficiency. Probably, in this case we have a considerable cardinal for the set Z_t, although the set F_t is not empty.

Tabla A4.3: Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared. Population: Simh.

	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 75$		$n = 100$		$n = 125$		$n = 150$	
$\widehat{F}_{HT}$	0.0066	1	0.0018	1	0.0036	1	0.0034	1
$\widehat{F}_{D}$	0.0075	0.5584	0.0016	0.5648	0.0027	0.5809	0.0041	0.5616
$\widehat{F}_{R}$	0.0078	1.2980	0.0018	1.2368	0.0031	1.2648	0.0041	1.2889
$\widehat{F}_{CD}$	0.3645	9.9689	0.3708	13.4606	0.3662	16.9590	0.3651	19.7880
$\widehat{F}_{RKM}$	0.0080	0.3990	0.0055	0.3822	0.0038	0.3990	0.0044	0.3789
$\widehat{F}_{CAL}$	0.0059	0.9369	0.0012	0.8205	0.0016	0.8794	0.0037	0.8845
$\widehat{F}_{CAL3}$	0.0028	0.3639	0.0015	0.3642	0.0008	0.3749	0.0037	0.3545
$\widehat{F}_{CALMAX}$	0.0012	0.2112	0.0008	0.1961	0.0016	0.1924	0.0009	0.1862
$\widehat{F}_{CALOPT}$	0.0030	0.1800	0.0012	0.1591	0.0006	0.1555	0.0010	0.1441
$\widehat{F}_{CALNEWOPT}$	0.0030	0.1800	0.0012	0.1591	0.0006	0.1555	0.0010	0.1441
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 175$		$n = 200$		$n = 250$		$n = 300$	
$\widehat{F}_{HT}$	0.0025	1	0.0037	1	0.0026	1	0.0015	1
$\widehat{F}_{D}$	0.0025	0.5626	0.0032	0.5822	0.0013	0.5593	0.0011	0.5653
$\widehat{F}_{R}$	0.0023	1.3001	0.0044	1.2704	0.0021	1.3077	0.0009	1.2659
$\widehat{F}_{CD}$	0.3653	23.6861	0.3743	26.3930	0.3659	32.5335	0.3632	38.7995
$\widehat{F}_{RKM}$	0.0040	0.3886	0.0024	0.3926	0.0020	0.3770	0.0011	0.3891
$\widehat{F}_{CAL}$	0.0021	0.8935	0.0048	0.8732	0.0011	0.8603	0.0012	0.8778
$\widehat{F}_{CAL3}$	0.0020	0.3780	0.0022	0.3721	0.0011	0.3480	0.0007	0.3598
$\widehat{F}_{CALMAX}$	0.0007	0.1920	0.0008	0.1853	0.0008	0.1758	0.0006	0.1725
$\widehat{F}_{CALOPT}$	0.0003	0.1486	0.0007	0.1438	0.0008	0.1303	0.0005	0.1313
$\widehat{F}_{CALNEWOPT}$	0.0003	0.1486	0.0007	0.1438	0.0008	0.1303	0.0005	0.1313

For the Simser population, Tables A4.5 and A4.6 provide the results of the simulation study. In this

Tabla A4.4: Average dimension(MD), variance dimension (VD) and comparison of execution time (RT) of the estimators $\widehat{F}_{CALOPT}$ and $\widehat{F}_{CALNEWOPT}$. Population: Simh.

	$n = 75$					$n = 100$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	2.355	0.543	1.395	0.489	0.557	2.471	0.527	1.489	0.500	0.780
t_2	4.168	0.529	1.44	0.584	0.400	4.313	0.503	1.597	0.655	0.299
t_3	5.18	0.560	1.55	0.639	0.710	5.341	0.522	1.739	0.676	0.541
t_4	6.289	0.616	1.73	0.697	0.364	6.47	0.551	1.91	0.696	0.335
t_5	8.552	0.657	2.277	0.739	0.519	8.769	0.533	2.471	0.675	0.343
t_6	10.509	0.702	2.416	0.795	0.504	10.78	0.633	2.681	0.811	0.327
t_7	12.514	0.863	2.831	0.998	0.367	12.833	0.769	3.162	0.956	0.388
t_8	14.475	0.822	2.923	0.956	0.242	14.807	0.645	3.239	0.877	0.351
t_9	15.664	0.928	3.1	1.0213	0.359	16.087	0.811	3.5	0.944	0.294
t_{10}	16.71	0.978	3.375	1.143	0.249	17.163	0.923	3.827	1.140	0.273
t_{11}	18.704	0.969	3.519	1.187	0.337	19.178	0.750	3.934	1.163	0.294
	$n = 125$					$n = 150$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	2.605	0.493	1.608	0.488	0.620	2.652	0.477	1.651	0.477	0.516
t_2	4.392	0.493	1.695	0.671	0.610	4.471	0.501	1.841	0.692	0.291
t_3	5.414	0.500	1.852	0.709	0.554	5.492	0.500	1.981	0.704	0.627
t_4	6.58	0.5018	2.041	0.714	0.488	6.673	0.469	2.224	0.687	0.659
t_5	8.883	0.456	2.64	0.628	0.339	8.961	0.414	2.775	0.573	0.409
t_6	10.928	0.541	2.865	0.734	0.488	11.058	0.517	3.033	0.717	0.377
t_7	12.988	0.658	3.415	0.892	0.343	13.092	0.621	3.578	0.876	0.342
t_8	14.95	0.586	3.412	0.824	0.208	15.09	0.520	3.657	0.812	0.275
t_9	16.277	0.753	3.698	0.936	0.308	16.394	0.708	3.862	0.896	0.304
t_{10}	17.42	0.809	4.127	1.096	0.353	17.526	0.810	4.293	1.031	0.241
t_{11}	19.438	0.641	4.333	1.089	0.262	19.532	0.584	4.564	1.087	0.278
	$n = 175$					$n = 200$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	2.719	0.450	1.719	0.450	0.602	2.763	0.426	1.763	0.426	0.603
t_2	4.507	0.500	1.929	0.699	0.711	4.581	0.494	1.991	0.688	0.613
t_3	5.551	0.498	2.056	0.712	0.450	5.588	0.492	2.167	0.701	0.277
t_4	6.714	0.454	2.288	0.657	0.466	6.776	0.417	2.431	0.634	0.674
t_5	9	0.364	2.858	0.508	0.372	9.022	0.360	2.901	0.485	0.366
t_6	11.062	0.500	3.088	0.685	0.473	11.117	0.494	3.211	0.727	0.328
t_7	13.178	0.617	3.74	0.848	0.440	13.214	0.631	3.846	0.835	0.355
t_8	15.136	0.542	3.765	0.795	0.299	15.191	0.495	3.897	0.754	0.310
t_9	16.474	0.694	4.003	0.859	0.303	16.576	0.667	4.209	0.830	0.326
t_{10}	17.632	0.806	4.514	0.981	0.340	17.726	0.799	4.652	1.021	0.260
t_{11}	19.617	0.524	4.747	1.021	0.268	19.652	0.523	4.891	1.009	0.227
	$n = 250$					$n = 300$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	2.832	0.374	1.832	0.374	0.702	2.894	0.308	1.894	0.308	0.763
t_2	4.659	0.474	2.135	0.691	0.522	4.714	0.452	2.29	0.668	0.519
t_3	5.675	0.469	2.319	0.668	0.537	5.758	0.429	2.517	0.588	0.610
t_4	6.858	0.349	2.596	0.563	0.480	6.93	0.255	2.712	0.479	0.336
t_5	9.069	0.304	3.01	0.387	0.408	9.115	0.346	3.09	0.385	0.369
t_6	11.178	0.465	3.319	0.685	0.453	11.242	0.467	3.392	0.702	0.297
t_7	13.312	0.635	4.061	0.844	0.349	13.405	0.621	4.209	0.817	0.297
t_8	15.268	0.490	4.056	0.702	0.272	15.357	0.506	4.211	0.699	0.325
t_9	16.67	0.649	4.365	0.786	0.330	16.797	0.664	4.582	0.787	0.322
t_{10}	17.897	0.766	4.914	0.980	0.310	18.041	0.735	5.179	0.904	0.304
t_{11}	19.777	0.419	5.157	0.902	0.323	19.821	0.386	5.317	0.892	0.294

case, Table A4.5 shows that $\widehat{F}_{CALOPT}(t)$ and $\widehat{F}_{CALNEWOPT}(t)$ do not present the same behavior and $\widehat{F}_{CALNEWOPT}(t)$ is the one with the least bias and the best efficiency of all the estimators included in the simulation study and it produce a considerable improvement in efficiency with respect to $\widehat{F}_{CALOPT}(t)$. For some of the estimators included in the simulation study, we can observe a worse efficiency than $\widehat{F}_{HT}(t)$, which may be caused by the absence of a linear relationship between y and x. Regarding dimension reduction analysis and efficiency in execution time, Table A4.6 shows that for all quantiles, the optimal vector of $\widehat{F}_{CALOPT}(t)$ increase its size when the sample size increases, especially in the larger quantiles, where we can observe very high dimensional optimal vectors. On the other hand, the dimension of the optimal vector for $\widehat{F}_{CALNEWOPT}(t)$ remains stable for all sample sizes and it always shows a value below 7 and in most cases its value is between 3 and 5. It represents a quite considerable reduction of the dimension that causes a quite remarkable improvement in execution, especially in the high quantiles and according to previous studies (Chauvet & Goga, 2022) this remarkable reduction allows $\widehat{F}_{CALNEWOPT}(t)$ to achieve an improvement in efficiency with respect to $\widehat{F}_{CALOPT}(t)$. In this case, the cardinal of the set Z_t is probably very high and it is similar to the cardinal of the set A_t, which implies that the set F_t has few elements.

6. Discussion and conclusions

In recent years, the calibration technique has attracted significant attention in survey sampling research and survey applications. The calibration method allows obtaining more reliable estimates for a finite population by incorporating auxiliary information available in the population.

In this article, we investigate whether the optimal estimator in the proposal Martínez et al. (2017) (that can be applied direclty in the estimation of quantile and poverty measures (Martínez et al., 2022)) based on the calibration method for estimating the distribution function can be improved by reducing the dimension of the optimal vector used in the calibration process. Working with a reduced number of variables may reduce numerical problems related to optimization procedures and also limit the presence of negative, very large and unstable calibration weights. To do this, we have theoretically established the conditions under which a reduction in the dimension of the optimal vector is possible and through an extensive simulation study we have verified how the new estimator $\widehat{F}_{CALNEWOPT}(t)$ can avoid the problems associated with a high-dimensional auxiliary data and allows to improve the execution time maintaining (DNase and Simh) or even improving the efficiency (Chauvet & Goga, 2022) (Simser). Therefore, the new proposal is a more reliable option when carrying out real analyzes where large population sizes can lead to high-dimensional optimal vectors for $\widehat{F}_{CALOPT}(t)$, while $\widehat{F}_{CALNEWOPT}(t)$ can lead to a considerable reduction in this optimal dimension.

Tabla A4.5: Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared. Population: Simser.

	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 75$		$n = 100$		$n = 125$		$n = 150$	
$\widehat{F}_{HT}$	0.0042	1	0.0049	1	0.0042	1	0.0032	1
$\widehat{F}_{D}$	0.0038	0.9963	0.0039	1.0008	0.0056	1.0022	0.0027	0.9999
$\widehat{F}_{R}$	0.0832	3.0296	0.0687	3.0130	0.0659	2.8977	0.0564	3.0535
$\widehat{F}_{CD}$	0.0305	0.8968	0.0215	0.9234	0.0196	0.9374	0.0117	0.9409
$\widehat{F}_{RKM}$	0.0177	1.0419	0.0106	1.0246	0.0196	1.0530	0.0110	1.0252
$\widehat{F}_{CAL}$	0.0271	1.5805	0.0199	1.5456	0.0274	1.5784	0.0201	1.5548
$\widehat{F}_{CAL3}$	0.0110	0.8375	0.0104	0.7839	0.0090	0.8138	0.0075	0.7860
$\widehat{F}_{CALMAX}$	0.0174	0.6580	0.0146	0.6495	0.0229	0.6857	0.0177	0.6580
$\widehat{F}_{CALOPT}$	0.0732	0.4007	0.0537	0.2894	0.0417	0.2307	0.0327	0.1799
$\widehat{F}_{CALNEWOPT}$	0.0028	0.1516	0.0017	0.1143	0.0015	0.0985	0.0016	0.0787
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 175$		$n = 200$		$n = 250$		$n = 300$	
$\widehat{F}_{HT}$	0.0029	1	0.0026	1	0.0011	1	0.0023	1
$\widehat{F}_{D}$	0.0037	1.0033	0.0022	0.9988	0.0013	1	0.0029	1.0016
$\widehat{F}_{R}$	0.0547	3.0697	0.0461	3.0468	0.0449	3.0664	0.0418	3.0488
$\widehat{F}_{CD}$	0.0094	0.9427	0.0139	0.9404	0.0127	0.9603	0.0121	0.9651
$\widehat{F}_{RKM}$	0.0129	1.0307	0.0072	1.0161	0.0079	1.0221	0.0090	1.0152
$\widehat{F}_{CAL}$	0.0215	1.5860	0.0143	1.4787	0.0160	1.5565	0.0169	1.5472
$\widehat{F}_{CAL3}$	0.0079	0.8379	0.0071	0.8280	0.0051	0.7523	0.0042	0.7902
$\widehat{F}_{CALMAX}$	0.0191	0.6812	0.0138	0.6426	0.0155	0.6591	0.0163	0.6747
$\widehat{F}_{CALOPT}$	0.0266	0.1529	0.0228	0.1248	0.0169	0.0900	0.0128	0.0733
$\widehat{F}_{CALNEWOPT}$	0.0010	0.0714	0.0007	0.0577	0.0007	0.0476	0.0004	0.0424

Tabla A4.6: Average dimension(MD), variance dimension (VD) and comparison of execution time (RT) of the estimators $\widehat{F}_{CALOPT}$ and $\widehat{F}_{CALNEWOPT}$. Population: Simser.

	$n = 75$					$n = 100$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	7.807	2.355	3.036	0.277	1.152	9.846	2.589	3.084	0.376	0.779
t_2	14.767	3.192	3.013	0.122	0.431	18.586	3.430	3.029	0.200	0.516
t_3	18.115	3.395	3.006	0.077	0.449	23.052	3.825	3.022	0.166	0.443
t_4	21.519	3.554	3.007	0.083	0.378	27.492	3.990	3.005	0.071	0.427
t_5	28.379	3.778	3.002	0.045	0.418	36.322	4.225	3.002	0.045	0.39
t_6	35.189	4.032	3.005	0.071	0.4	45.339	4.389	3.013	0.113	0.318
t_7	42.003	4.073	3.015	0.122	0.296	54.139	4.445	3.005	0.071	0.283
t_8	48.955	3.807	3	0	0.305	63.065	4.240	3.006	0.077	0.294
t_9	52.367	3.655	3.003	0.055	0.338	67.537	4.076	3.009	0.094	0.237
t_{10}	55.801	3.565	3.004	0.063	0.246	71.983	3.950	3.002	0.089	0.248
t_{11}	62.753	3.023	2.948	0.363	0.228	80.714	3.531	2.982	0.289	0.268
	$n = 125$					$n = 150$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	11.934	3.145	2.664	0.471	0.946	13.766	2.767	3.141	0.455	0.762
t_2	22.63	3.046	3.697	0.241	0.331	26.001	3.740	3.066	0.303	0.453
t_3	27.879	3.023	3.967	0.169	0.42	32.224	4.073	3.027	0.185	0.337
t_4	33.406	3.014	4.195	0.118	0.348	38.628	4.411	3.013	0.113	0.379
t_5	44.154	3.011	4.626	0.122	0.325	51.218	4.853	3.008	0.090	0.287
t_6	54.934	3.011	4.893	0.104	0.283	63.845	5.074	3.026	0.159	0.259
t_7	65.748	3.01	4.875	0.010	0.253	76.315	5.219	3.019	0.137	0.243
t_8	76.511	3.01	4.815	0.010	0.191	88.878	5.144	3.006	0.077	0.19
t_9	81.862	3.01	4.695	0.010	0.255	95.292	5.014	3.013	0.113	0.212
t_{10}	87.184	3.013	4.555	0.113	0.192	101.589	4.937	3.02	0.140	0.218
t_{11}	98.008	3	4.022	0.219	0.208	114.415	4.415	3.027	0.180	0.195
	$n = 175$					$n = 200$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	15.496	3.036	2.927	0.277	0.855	16.921	2.936	3.258	0.631	0.663
t_2	29.729	3.013	3.853	0.122	0.434	32.618	3.922	3.089	0.351	0.367
t_3	36.67	3.006	4.144	0.077	0.438	40.25	4.137	3.058	0.277	0.26
t_4	43.817	3.007	4.441	0.083	0.335	48.197	4.458	3.037	0.189	0.276
t_5	58.001	3.002	4.757	0.045	0.268	63.872	4.957	3.012	0.109	0.26
t_6	72.284	3.005	5.065	0.071	0.227	79.743	5.293	3.033	0.179	0.234
t_7	86.505	3.015	5.062	0.122	0.25	95.683	5.487	3.061	0.239	0.209
t_8	100.946	3	5.083	0	0.219	111.761	5.488	3.011	0.104	0.191
t_9	108.154	3.003	4.976	0.055	0.194	119.62	5.366	3.022	0.147	0.195
t_{10}	115.308	3.004	4.883	0.063	0.18	127.542	5.398	3.031	0.173	0.171
t_{11}	129.69	2.948	4.693	0.363	0.175	143.597	5.105	3.056	0.230	0.149
	$n = 250$					$n = 300$				
	MD		VD		RT	MD		VD		RT
	OPT	$NEWOPT$	OPT	$NEWOPT$		OPT	$NEWOPT$	OPT	$NEWOPT$	
t_1	20.073	3.438	3.111	0.787	0.552	22.612	3.533	3.117	0.860	0.646
t_2	38.419	3.173	4.275	0.491	0.294	43.526	3.193	4.048	0.484	0.381
t_3	47.492	3.091	4.653	0.342	0.29	53.736	3.131	4.393	0.387	0.323
t_4	57.048	3.054	5.007	0.226	0.288	64.659	3.072	4.745	0.259	0.301
t_5	75.652	3.015	5.438	0.122	0.25	85.713	3.026	5.212	0.177	0.206
t_6	94.572	3.058	5.786	0.234	0.204	107.232	3.081	5.639	0.273	0.191
t_7	113.331	3.058	5.855	0.234	0.186	128.379	3.094	5.858	0.292	0.177
t_8	132.281	3.024	5.856	0.153	0.174	149.69	3.035	5.987	0.184	0.15
t_9	141.666	3.035	5.778	0.184	0.156	160.366	3.036	6.071	0.186	0.136
t_{10}	150.953	3.043	5.768	0.203	0.14	171.005	3.067	6.076	0.250	0.135
t_{11}	169.593	3.083	5.433	0.276	0.124	192.424	3.096	5.887	0.295	0.132

Further research is needed regarding the dimension reduction on calibration for the distribution function as our study presents certain limitations. Our paper is restricted to a simple random sampling design. The determination of the optimal vector for calibration (and its dimension) can be extended relatively easily to the case of self-weighted samples (for example, stratified samples with proportional allocation). However, the case of sampling with unequal probabilities is more complex and the methodology to be used is not the same. In future research we try to extend the results of this paper from SRSWOR to complex sampling designs.

Another limitation of our work is that the estimator considered is based on a pseudo-variable g_k that assumes a linear relationship between variable y and the covariates. The selection of the optimal auxiliary vector for the estimators based on a non-linear model should be considered in future studies.

Financial disclosure

This study was partially supported by Ministerio de Educación y Ciencia (PID2019-106861RB-I00, Spain), IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033 and FEDER/Junta de Andalucía-Consejería de Transformación Econiómica, Industria, Conocimiento y Universidades (FQM170-UGR20).

Conflict of interest

The authors declare no potential conflict of interests.

1. Supplementary Cases for Section 3

1.1. Dimension reduction of the optimal auxiliary vector for $\mathbf{t} = \mathbf{y_{max}}$

If we consider $t = y_{max}$ where

$$y_{max} = \max_{k \in U} y_k$$

the set A_t is given by

$$A_t = \{g_k : k \in U; y_k \leq t\} = \{a_1, a_2, \ldots, a_M\} = A_M \tag{26}$$

with $a_1 < a_2 < \cdots < a_M$ and M denotes the total number of different values that the pseudo variable g can take in the population U. As a consequence, $B_t = \emptyset$, the optimal dimension $P = M$ is the highest value and

the optimal vector is given by:

$$\mathbf{t_{OP}}(t) = (a_1, a_2, \ldots, a_M).$$

Our purpose is to analyze the possibility of obtaining the minimum value of (4.14) by means of a lower-dimensional auxiliary vector. Firstly, if we consider the value $t = y_{max}$, we have:

$$K_t(a_j) = \sum_{k \in U} \Delta(a_j - g_k)\Delta(y_{max} - y_k) = N \cdot F_g(a_j) \quad j = 1, \ldots, M$$

and where we set a_0 so that $F_g(a_0) = 0$ and $K_t(a_0)$.
The value of $Q_t(\gamma)$ at $\gamma = \mathbf{t_{OP}}(t)$ is given by:

$$Q_t(\mathbf{t_{OP}}(t)) = Q_t(a_1, a_2 \ldots, a_M) = 2NF_y(y_{max}) \cdot K_t(a_M) - \sum_{j=1}^{M} \frac{\big(K_t(a_j) - K_t(a_{j-1})\big)^2}{(F_g(a_j) - F_g(a_{j-1}))} - \big(K_t(a_M)\big)^2 =$$

$$= 2NF_y(y_{max}) \cdot N \cdot F_g(a_M) - \sum_{j=1}^{M} \frac{\big(F_g(a_j) - F_g(a_{j-1})\big)^2}{(F_g(a_j) - F_g(a_{j-1}))} - \big(N \cdot F_g(a_M)\big)^2 =$$

$$= 2NF_y(y_{max}) \cdot N \cdot F_g(a_M) - \sum_{j=1}^{M} N^2 \cdot (F_g(a_j) - F_g(a_{j-1})) - \big(N \cdot F_g(a_M)\big)^2 = 2N^2 F_y(y_{max}) \cdot F_g(a_M) - \big(N \cdot F_g(a_M)\big)^2.$$

Since $F_y(y_{max}) = F_g(a_M) = 1$, it is clear that $Q_t(\mathbf{t_{OP}}(t)) = 0$ and consequently the minimum value of $Q_t(\gamma)$ for y_{max} is equal to 0.

On the other hand, if we consider $\gamma = (a_M)$ then

$$Q_t(a_M) = 2NF_y(y_{max}) \cdot K_t(a_M) - \frac{\big(K_t(a_M)\big)^2}{(F_g(a_j))} - \big(K_t(a_M)\big)^2 =$$

$$Q_t(a_M) = 2N^2 F_y(y_{max}) \cdot F_g(a_M) - 2N^2 \big(F_g(a_M)\big)^2 = 0.$$

Thus, with the auxiliary vector $\gamma = (a_M)$ the minimum value of $Q_t(\gamma)$ is reached and the optimal dimension can be reducted from M to 1. With the auxiliary vector $\gamma = (a_M)$, the resulting calibration constraint is given by:

$$1 = F_g(a_M) = \frac{1}{N} \sum_{k \in s} \omega_k \Delta(a_M - g_k) = \frac{1}{N} \sum_{k \in s} \omega_k \tag{27}$$

Under simple random sampling without replacement, the minimization of (4.5) subject to the condition (27) results in $d_k = \omega_k$ since the basic weights d_k associated with the simple random sampling without replacement satisfy the condition (27). Therefore, the minimum value of $Q_t(\gamma)$ for y_{max} is equal to 0.

1.2. Dimension reduction of the optimal auxiliary vector when $\mathbf{D_t = \emptyset; Z_t = \emptyset}$ and $\mathbf{F_t = A_t = A_M}$

If we consider the case where $D_t = \emptyset$; $Z_t = \emptyset$ and $F_t = A_t = A_M$ there is not reduction in the optimal auxiliary vector $\mathbf{t_{OP}}(t)$. To see it, it is clear that $q_i^t \neq 0 \ \forall a_i \in A_M$ and consequently:

$$K_t(a_i) = \sum_{inU} \Delta(a_i - g_k)\Delta(t - y_k) = N \cdot F_g(a_i) - \sum_{j=1}^{i} q_j^t \text{ for } i = 1, 2, \ldots, M.$$

Specifically, for $i = M$ we have:

$$K_t(a_M) = N \cdot F_g(a_M) - \sum_{j=1}^{i} q_j^t = NF_y(t).$$

The minimum value of $Q_t(\gamma)$ is reached at $\gamma = \mathbf{t_{OP}}(t)$ and is given by:

$$Q_t(\mathbf{t_{OP}}(t)) = 2NF_y(t) \cdot K_t(a_M) - \sum_{j=1}^{M} \frac{\left(K_t(a_j) - K_t(a_{j-1})\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \left(K_t(a_M)\right)^2.$$

Since $(K_t(a_j) - K_t(a_{j-1}) = N \cdot F_g(a_j) - N \cdot F_g(a_{j-1}) - q_j^t$, $Q_t(\mathbf{t_{OP}}(t))$ takes the following expression:

$$Q_t(\mathbf{t_{OP}}(t)) = (NF_y(t))^2 - \sum_{j=1}^{M} \frac{\left(N \cdot F_g(a_j) - N \cdot F_g(a_{j-1}) - q_j^t\right)^2}{F_g(a_j) - F_g(a_{j-1})} =$$

$$(NF_y(t))^2 - N^2 \cdot \sum_{j=1}^{M} \left(F_g(a_j) - F_g(a_{j-1})\right) + 2N \cdot \sum_{j=1}^{M} q_j^t - \sum_{j=1}^{M} \frac{(q_j^t)^2}{F_g(a_j) - F_g(a_{j-1})} =$$

$$(NF_y(t))^2 - N^2 \cdot F_g(a_M) + 2N \cdot \sum_{j=1}^{M} q_j^t - \sum_{j=1}^{M} \frac{(q_j^t)^2}{F_g(a_j) - F_g(a_{j-1})} = (NF_y(t))^2 - N^2 + 2N \cdot \sum_{j=1}^{M} q_j^t - \sum_{j=1}^{M} \frac{(q_j^t)^2}{F_g(a_j) - F_g(a_{j-1})}. \tag{28}$$

If we consider an auxiliary vector where we delete some value $a_i \neq a_M$, i.e $\gamma = (a_1, \ldots a_{i-1}, a_{i+1}, \ldots, a_M)$, we can obtain in a similar way that

$$Q_t(\gamma) = 2NF_y(t) \cdot K_t(a_M) - \sum_{j=1}^{i-1} \frac{\left(K_t(a_j) - K_t(a_{j-1})\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \sum_{j=i+1}^{M} \frac{\left(K_t(a_j) - K_t(a_{j-1})\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \frac{\left(K_t(a_{i+1}) - K_t(a_{i-1})\right)^2}{F_g(a_{i+1}) - F_g(a_{i-1})} - \left(K_t(a_M)\right)$$

$$= (NF_y(t))^2 - N^2 + 2N \cdot \sum_{\substack{j=1 \\ j \neq i, i+1}}^{M} q_j^t + 2N(q_i^t + q_{i+1}^t) - \sum_{\substack{j=1 \\ j \neq i, i+1}}^{M} \frac{(q_j^t)^2}{F_g(a_j) - F_g(a_{j-1})} - \frac{(q_i^t + q_{i+1}^t)^2}{F_g(a_{i+1}) - F_g(a_{i-1})}.$$

As a consequence, $Q_t(\mathbf{t_{OP}}(t)) - Q_t(\gamma)$ is given by

$$Q_t(\mathbf{t_{OP}}(t)) - Q_t(\gamma) = -\frac{(q_i^t)^2}{F_g(a_i) - F_g(a_{i-1})} - \frac{(q_{i+1}^t)^2}{F_g(a_{i+1}) - F_g(a_i)} + \frac{(q_i^t + q_{i+1}^t)^2}{F_g(a_{i+1}) - F_g(a_{i-1})} =$$

$$= -\frac{(q_i^t)^2\big(F_g(a_{i+1}) - F_g(a_i)\big)}{\big(F_g(a_i) - F_g(a_{i-1})\big)\big((F_g(a_{i+1}) - F_g(a_{i-1})\big)} - \frac{(q_{i+1}^t)^2\big(F_g(a_i) - F_g(a_{i-1})\big)}{\big(F_g(a_{i+1}) - F_g(a_{i-1})\big)\big((F_g(a_{i+1}) - F_g(a_i)\big)} + \frac{2q_i^t \cdot q_{i+1}^t}{F_g(a_{i+1}) - F_g(a_{i-1})} =$$

$$= \Gamma \cdot \Big[-(q_i^t)^2\big(F_g(a_{i+1}) - F_g(a_i)\big)^2 - (q_{i+1}^t)^2\big(F_g(a_i) - F_g(a_{i-1})\big)^2 + 2q_i^t \cdot q_{i+1}^t\big(F_g(a_{i+1}) - F_g(a_i)\big)\big(F_g(a_i) - F_g(a_{i-1})\big)\Big] < 0$$

with

$$\Gamma = \frac{1}{\big(F_g(a_i) - F_g(a_{i-1})\big)\big(F_g(a_{i+1}) - F_g(a_i)\big)\big(F_g(a_{i+1}) - F_g(a_{i-1})\big)}.$$

Consequently, when deleting some a_i, $Q_t(\mathbf{t_{OP}}(t)) < Q_t(\gamma)$.

If we delete the value a_M, i.e, we consider the auxiliary vector $\gamma = (a_1, \ldots, a_{M-1})$, $Q_t(\gamma)$ takes the following expression:

$$Q_t(\gamma) = 2NF_y(t) \cdot K_t(a_{M-1}) - \sum_{j=1}^{M-1} \frac{\big(K_t(a_j - K_t(a_{j-1}))\big)^2}{F_g(a_j) - F_g(a_{j-1})} - (K_t(a_{M-1}))^2$$

On the other hand

$$Q_t(\mathbf{t_{OP}}(t)) = (NF_y(t))^2 - \sum_{j=1}^{M} \frac{\big(K_t(a_j - K_t(a_{j-1}))\big)^2}{F_g(a_j) - F_g(a_{j-1})}$$

and it easy to see that

$$Q_t(\mathbf{t_{OP}}(t)) - Q_t(\gamma) = (NF_y(t) - K_t(a_{M-1}))^2 - \frac{(NF_y(t) - K_t(a_{M-1}))^2}{F_g(a_M) - F_g(a_{M-1})} < 0.$$

Therefore, when we delete a_M; $Q_t(\mathbf{t_{OP}}(t)) < Q_t(\gamma)$.

Thus, if $Z_t = \emptyset$ and $F_t = A_t$ there is not a reduction in the auxiliary vector to reach the minimum of $Q_t(\gamma)$.

1.3. Dimension reduction of the optimal auxiliary vector for p_i; $i \in \{2, \ldots, l_t\}$ when $\mathbf{D_t} \neq \emptyset$; $\mathbf{D_t} = \mathbf{A_M}$ and $\mathbf{B_t} \neq \emptyset$

Under the assumptions $\mathbf{D_t} \neq \emptyset$; $\mathbf{D_t} = \mathbf{A_M}$ and $\mathbf{B_t} \neq \emptyset$, if we consider p_i with $i = 2, \ldots, l_t$, it is clear that $p_i > p_{(i-1)}$ and $f_{p_i}^t > f_{p_{(i-1)}}^t$. Moreover, because the value $b_{f_{p_i}}^t$ exists, this implies that $f_{p_i}^t > f_{p_{(i-1)}}^t + 1$.

If we suppose that $p_i = p_{(i-1)} + 1$, it is clear that $f^t_{p_i} = f^t_{(p_{(i-1)}+1)}$ and due to the value $b^t_{f_{p_i}}$ exists, the set

$$U_{p_{(i-1)}+1} = U_{p_i} = \{l \in U : a_{f^t_{p_{(i-1)}}} < g_l < a_{f^t_{p_{(i-1)}+1}}\} = \{l \in U : a_{f^t_{p_{(i-1)}}} < g_l < a_{f^t_{p_i}}\} \neq \emptyset$$

As a consequence, we have:

$$\{a_{f^t_{(p_{(i-1)}+1)}}, \ldots, a_{(f^t_{p_i}-1)}\} \subseteq D_t$$

and $b^t_{f_{p_i}} = a_{(f^t_{p_i}-1)}$ and there is not a possible reduction in the dimension.

On the contrary, if we suppose that $p_i > p_{(i-1)} + 1$, then $f^t_{p_i} > f^t_{(p_{(i-1)}+1)}$ and there is a integer $z \geq 1$ such that $p_i = p_{(i-1)} + 1 + z$. For all $j = 1, \ldots, z$, the value $b_{f^t_{p_{(i-1)}+j}}$ does not exist and the set $U_{p_{(i-1)}+j} = \emptyset$. As in the previous case (case p_1), we have:

$$a_{f^t_{(p_{(i-1)}+j)}} = a_{(f^t_{p_{(i-1)}}+j)}, \quad j = 1, \ldots, z.$$

Thus, if $p_i > p_{(i-1)} + 1$, we have:

$$\{a_{(f^t_{p_{(i-1)}}+1)}, \ldots, a_{(f^t_{p_{(i-1)}}+p_i-p_{(i-1)}-1)}\} \subseteq A_t.$$

Then, if we define the following sets:

$$A_{p_i} = \{a_{(f^t_{p_{(i-1)}}+1)}, \ldots, a_{(f^t_{p_{(i-1)}}+p_i-p_{(i-1)}-1)}\}$$

$$Z_{p_i} = \{a_i \in A_{p_i} : q^t_i = 0\}$$

and

$$F_{p_i} = \{a_i \in A_{p_i} : 0 < q^t_i < r_i\}$$

we can proof in a similar way to the previous case (case p_1) that if $Z_{p_i} = A_{p_i}$ or $Z_{p_i} \neq A_{p_i}$ but $F_{p_i} \neq A_{p_i}$ there is a possible reduction in the dimension of the auxiliary vector $\mathbf{t_{OP}}(t)$. If $F_{p_i} = A_{p_i}$ there is no possible dimension reduction.

Finally, if we suppose that $p_{l_t} = M_t$ then the value $b_{f^t_{M_t}}$ exists and analogously to previous cases, there is no reduction between the points $b_{f^t_{M_t}}$ and $a_{f^t_{M_t}}$.

On the other hand, if we suppose that $p_{l_t} < M_t$ then for $h = p_{l_t} + 1, p_{l_t} + 2, \ldots, M_t$ the corresponding

value $b_{f_h^t}$ does not exist and therefore the sets $U_h = \emptyset$. As a consequence, we have:

$$a_{f^t_{(p_{l_t}+1)}} = a_{(f^t_{p_{l_t}}+1)}, \ldots, a_{f^t_{M_t}} = a_{(f^t_{p_{l_t}}+M_t-p_{l_t})}$$

If we denote by

$$A_{M_t} = \{a_{(f^t_{p_{l_t}}+1)}, \ldots, a_{(f^t_{p_{l_t}}+M_t-p_{l_t})}\} \subseteq A_t$$

$$Z_{M_t} = \{a_i \in A_{M_t} : q_i^t = 0\}$$

$$F_{M_t} = \{a_i \in A_{M_t} : 0 < q_i^t < r_i\}$$

then, we can reduce the dimension of the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ if $Z_{M_t} = A_{M_t}$ or if $Z_{M_t} \neq A_{M_t}$ but $F_{M_t} \neq A_{M_t}$. If $F_{M_t} = A_{M_t}$ there is no possible dimension reduction.

References

[1] Arcos, A., Martínez, S., Rueda, M., & Martínez, H. (2017). Distribution function estimates from dual frame context, *Journal of Computational and Applied Mathematics*, 318, 242-252.

[2] Breidt, F. J., Opsomer, J. D., Johnson, A. A., & Ranalli, M. G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology*, 33(1), 35.

[3] Cardot, H., Goga, C., & Shehzad, M. A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 243-260.

[4] Chambers, R.L., & Clark, RG. (2008). Adaptive calibration for prediction of finite population totals. *Survey Methodology*, 34(2), 163-172.

[5] Chambers, R. L., & Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597-604.

[6] Chen, J., & Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 1223-1239.

[7] Devaud, D., & Tillé, Y. (2019). Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem. *Test*, 28(4), 1033-1065.

[8] Deville,J.C., & Särndal, C.E.(1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87(418), 376-382.

[9] Dickens, R., & Manning, A. (2004). Has the national minimum wage reduced UK wage inequality?. *Journal of the Royal Statistical Society: Series A*, 167(4), 613-626.

[10] Estevao, V. M., & Särndal, C. E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2), 127-147.

[11] Guggemos, F., & Tille, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference*, 140(11), 3199-3212.

[12] Harms, T., & Duchesne, P. (2006). On calibration estimation for quantiles, *Survey Methodology*, 32, 37-52.

[13] Kovacevic, M. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. In *Proceedings of the Survey Methods Section, Statistical Society of Canada* (Vol. 47, pp. 139-144).

[14] Machin, S., Manning, A., & Rahman, L. (2003). Where the minimum wage bites hard: introduction of minimum wages to a low wage sector. *Journal of the European Economic Association*, 1(1), 154-180.

[15] Martínez, S., Rueda, M., Arcos, A., & Martínez, H. (2010). Optimum calibration points estimating distribution functions. *Journal of Computational and Applied Mathematics*, 233(9), 2265-2277.

[16] Martínez, S., Rueda, M., Arcos, A., Martínez, H., & Sánchez-Borrego, I. (2011). Post-stratified calibration method for estimating quantiles. *Computational Statistics and Data Analysis*, 55(1), 838-851.

[17] Martínez, S., Rueda, M., Martínez, H., & Arcos, A. (2015). Determining P optimum calibration points to construct calibration estimators of the distribution function. *Journal of Computational and Applied Mathematics*, 275, 281-293.

[18] Martínez, S., Rueda, M., Martínez, H., & Arcos, A. (2017). Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function. *Journal of Computational and Applied Mathematics*, 318, 444-459.

[19] Martínez, S., Rueda, M., & Illescas, M. (2020). The optimization problem of quantile and poverty measures estimation based on calibration. *Journal of Computational and Applied Mathematics*, 113054.

[20] Mayor-Gallego, J. A., Moreno-Rebollo, J. L., & Jiménez-Gamero, M. D. (2019). Estimation of the finite population distribution function using a global penalized calibration method. *AStA Advances in Statistical Analysis*, 103(1), 1-35.

[21] McConville, K. S., Breidt, F. J., Lee, T., & Moisen, G. G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5(2), 131-158.

[22] Nickell, S. (2004). Poverty and worklessness in Britain. *The Economic Journal*, 114(494), C1-C25.

[23] Rao, J. N. K. and Kovar, J. G. & Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information, *Biometrika*, 77(2), 365-375.

[24] Rota, B. J. (2017). Variance estimation in two-step calibration for nonresponse adjustment. *South African Statistical Journal*, 51(2), 361-374.

[25] Rueda, M. D. M. (2019). Comments on: Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. *Test*, 28(4), 1077-1081.

[26] Rueda, M., Martínez, S., Martínez, H., & Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137(2), 435-448.

[27] Sedransk, N., & Sedransk, J. (1979). Distinguishing among distributions using data from complex sample designs. *Journal of the American Statistical Association*, 74(368), 754-760.

[28] Nascimento Silva, P. L. D., & Skinner, C. J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23(1), 23-32.

[29] Singh, S. (2001). Generalized calibration approach for estimating variance in survey sampling. *Annals of the Institute of Statistical Mathematics*, 53(2), 404-417.

[30] Singh, H. P., Singh, S., & Kozak, M. (2008). A family of estimators of finite-population distribution function using auxiliary information. *Acta applicandae mathematicae*, 104(2), 115-130.

[31] Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90(4), 937-951.