Boletín de Estadística e Investigación Operativa
Vol. 36, No. 2, Julio 2020, pp. 153-171

Estadística Oficial

Sample design to monitor COVID-19 disease

Domingo Morales
Universidad Miguel Hernández de Elche, España
✉ d.morales@umh.es

María José Lombardía
CITIC, Universidade da Coruña, España
✉ maria.jose.lombardia@udc.es

Ricardo Fraiman
Universidad de la República, Uruguay
✉ fraimanricardo@gmail.com

Juan Antonio Cuesta Albertos
Universidad de Cantabria, España
✉ juan.cuesta@unican.es

Abstract

This paper contains some proposals for sampling oriented to the weekly estimation of the real proportion of individuals that have been infected (present or past infection) by SARS-COV-2. During the months of May to July 2020, Instituto de Salud Carlos III in collaboration with the Instituto Nacional de Estadística carried out three monthly surveys to estimate the number of people with virus antibodies in each province. Complementing the results of these surveys, increasing their periodicity to weeks, would be of great help for health decision-making.

Keywords: COVID-19, Prevalence estimation, Sequential sampling, Survey sampling.
AMS Subject classifications: 62D05.

1. Introducción

Este artículo introduce y discute algunas propuestas para la obtención de estimaciones fiables, con periodicidad semanal, de la prevalencia de anticuerpos anti SARS-COV-2 en zonas de alta densidad de población. Ello permite implicar en el estudio a los ayuntamientos con mayores recursos económicos y aumentar la periodicidad de las estimaciones. Se pretende complementar el estudio nacional de sero-epidemiología de la infección por SARS-COV-2, realizado en España por el Instituto de Salud Carlos III en colaboración del Instituto Nacional de Estadística (INE).

En términos más coloquiales, se quiere estimar la proporción de personas que tienen o han pasado la enfermedad COVID-19, lo que indicaremos abreviadamente en este artículo por prevalencia. Esto incluye, tanto los casos activos en un determinado momento, como a las personas que han pasado la infección de un modo extremadamente leve o, incluso, de un modo asintomático. Con el objetivo de rebajar los costes del trabajo de campo, proponemos limitar el estudio a las capitales de provincia y a un número determinado de ciudades grandes, limitado por la capacidad de muestreo de que se disponga.

El protocolo para contactar a los participantes es el mismo que en la encuesta ENE-COVID19, requiriendo un contacto telefónico previo. A cada participante se le solicita una extracción de fluidos de nasofaringe para hacer las pruebas PCR de detección de rastros del virus y una muestra de sangre extraída por venopunción para hacer el análisis serológico para la detección de anticuerpos IgM/IgG anti SARS-CoV-2 mediante métodos de alto rendimiento tipo ELISA. Además, se hace un test rápido de anticuerpos por digitopunción y se cumplimenta un cuestionario médico. La toma de datos se realiza en centros de salud con cita previa, salvo en casos excepcionales con visitas a domicilios. De ambas pruebas se pueden obtener estimaciones de la prevalencia. La prueba ELISA tiene mayor precisión en la detección de anticuerpos, pero tiene el problema de que hay mucha gente reacia a la venopunción. La prueba rápida tiene un menor grado de precisión, pero no genera rechazo.

Este documento no pretende ser una propuesta cerrada. Su finalidad es servir como punto de partida para avanzar en el objetivo planteado. Así, nuestros objetivos específicos para España serían los siguientes:

- Hacer un estudio limitado a las ciudades que son capitales de provincia o equivalentes. Por ejemplo, 60 ciudades. Para cada ciudad habrá un equipo de trabajo que hará un estudio independiente.

- Publicar estimaciones semanales basadas en muestras pequeñas. Por ejemplo, realizando al menos $n = 250$ tests. Al contrario del muestreo ENE-COVID19, que usa una muestra panel para estudiar la evolución de la prevalencia, el énfasis se pone en la estimación precisa en cada periodo de

tiempo. Por tal motivo, se extraen muestras semanales independientes y se aconseja usar estimadores suavizados usando datos de semanas anteriores.

- Publicar mensualmente estimaciones más precisas que las semanales, usando las muestras de 4 semanas consecutivas.

Para ello en la Sección 2 se presentan las ideas básicas y algunas sugerencias con ejemplos. Las secciones 3, 4 y 5 proporcionan los desarrollos matemáticos necesarios para la implantación de los diseños muestrales sugeridos combinando las ideas mostradas en la Sección 2. La redacción de estos apartados es muy técnica y va dirigida a especialistas en muestreo e inferencia en poblaciones finitas.

2. Elementos del estudio

En esta sección se dan las ideas básicas y la propuesta de actuación que se particularizarán en los diseños mostrados en las siguientes secciones. Nuestra propuesta tiene dos fases, aunque podría ponerse en marcha solamente la primera. Ello dependerá, entre otros, de factores presupuestarios y de tiempo de ejecución. En la primera fase se realizará un muestreo bietápico, de modo similar a la primera ronda de la encuesta ENE-COVID19. Se propone entonces un diseño muestral similar al de la Encuesta de Población Activa (EPA) del INE. A continuación, se pasa a una segunda fase secuencial para profundizar en el entorno de las personas que hayan resultado positivas en los test de anticuerpos de la fase inicial.

2.1. Toma de datos

La *entrevista* a los encuestado consite en:

1. Realizar un test rápido por digitopunción, un test ELISA por venopunción y un test PCR en fluidos de nasofaringe para SARS-COV-2.

2. Cumplimentar un cuestionario que incluya variables de información auxiliar básica relativa al encuestado, al hogar y a datos individuales de tipo epidemiológico, de salud, de hábitos de transporte y viaje, de ámbito laboral y socioeconómicos. Se recomienda cubrir el mismo cuestionario que la encuesta ENE-COVID19. Ello permitirá construir estimadores indirectos basados en el diseño o basados en modelos.

2.2. Fase I: Estudio no secuencial

Se reparte (estratifica) la población en unos pocos estratos. Lo estadísticamente más eficiente sería agrupar por zonas geográficas que tengan alguna relación con la intensidad de la epidemia. Alternativamente, la agrupación se

puede hacer por barrios o conjuntos de distritos municipales. Idealmente, la población dentro de cada estrato debe ser homogénea respecto de la variable de interés, pero los estratos han de ser heterogéneos entre sí. Las dos etapas de muestreo de la fase inicial son:

Etapa 1. Cada estrato se divide en conglomerados, que son subzonas muy pequeñas dentro de los estratos. Por lo tanto, típicamente, los estratos están divididos en muchos conglomerados. Lo ideal es que los conglomerados sean "fotocopias" reducidas del estrato al que pertenecen; es decir, la heterogeneidad del estrato debe reproducirse en todos los conglomerados, que, por lo tanto, deben ser similares entre sí. Se selecciona una muestra aleatoria simple de conglomerados en cada estrato.

Etapa 2. Se selecciona una muestra de viviendas dentro de cada conglomerado seleccionado. Se contacta telefónicamente y se cita a todos los miembros de las viviendas seleccionadas en centros de salud para hacer las pruebas de antígenos y/o de anticuerpos que se hayan previsto.

La vivienda (entendida como vivienda familiar principal) no cubre toda la población, pues supone dejar fuera del marco poblacional a la población que habita en hogares colectivos (residencias de ancianos, centros de discapacitados, hospitales de crónicos, establecimientos penitenciarios, cuarteles, conventos, etc.). Previsiblemente, esto introduciría una estimación a la baja de la proporción real. Pero, para fijar ideas, preferimos continuar con este planteamiento. Esto podrá resolverse haciendo un estudio específico para la población que habita en hogares colectivos.

Como ejemplo ilustrativo, para una ciudad dada, por ejemplo, Madrid, el diseño de la Fase I podría ser el siguiente:

Etapa 1. Repartir sus 22 distritos en 5 estratos. Para cada estrato formamos un equipo que haga independientemente el trabajo de campo (contacto telefónico, toma de datos en centros de salud o domicilios) y de digitalización.

Los conglomerados pueden coincidir con las secciones censales (zonas geográficamente poco extensas con 2000 habitantes por término medio). De cada estrato seleccionamos al azar 5 conglomerados. Se dispone de 5 telefonistas con formación sanitaria, de modo que a cada telefonista se le asigna el contacto con los hogares de un conglomerado.

Etapa 2. De cada conglomerado, seleccionamos 10 hogares con muestreo aleatorio simple sin reemplazamiento. El telefonista contacta con las viviendas seleccionadas y da las citas para que acudan a los centros de salud para hacer los tests serológicos y cumplimentar el cuestionario. Se hacen varios intentos de contacto con objeto de reducir la no respuesta.

De acuerdo con el ejemplo anterior, cada semana se obtendrá una muestra de al menos 250 participantes ($= 5 \cdot 5 \cdot 10$) en la ciudad de Madrid.

2.3. Fase II: Estudio secuencial

En la Fase II se identificarán y evaluarán todas las personas que hubieran estado en contacto con cada infectado (PCR positivo) de la primera fase durante un plazo a determinar. Por ejemplo, en los 15 días inmediatamente anteriores a la realización del test.

Es obvio que la proporción de positivos identificados con este procedimiento es un estimador (muy) sesgado al alza de la verdadera proporción de infectados. No obstante, las ideas del *clustering sampling* (ver [6] y [7]) llevan a la conclusión de que si se utiliza una media adecuadamente ponderada de los resultados obtenidos, no solo se obtiene un estimador insesgado, sino que, adicionalmente, se reduce considerablemente el error cuadrático medio de la media muestral. Además, este estimador mejora apreciablemente al que se obtendría utilizando un muestreo tradicional con el mismo tamaño muestral, sobre todo en las zonas con incidencia baja del virus.

Este procedimiento se basa, en cierto modo, en una de las ideas básicas empleadas en la actualidad para controlar la difusión del virus. Se trata de identificar con la mayor celeridad posible a todos los posibles afectados. La diferencia estriba en el objetivo buscado:

- Si el objetivo es construir una gran base de datos de infectados, presentes y pasados, para, por ejemplo, planificar una suavización de las medidas de distancia social, la segunda fase debe reiterarse hasta que llegue un momento en el que todos los contactos estén sanos partiendo también de un tamaño de la muestra inicial bastante mayor.

- Si el objetivo es la estimación periódica de la prevalencia, la iteración deberá parar después de cierto número de repeticiones de la Fase II (quizás no más de cuatro o cinco), con la finalidad de repetir la Fase I para empezar con otra muestra que permita obtener información de otros entornos y, con ello, reducir el error del muestreo.

Una limitación del procedimiento propuesto es que el tamaño muestral definitivo no está fijado a priori y, por lo tanto, los costes de la implementación estarán indeterminados. Una posibilidad (que, desde luego, limita las bondades del procedimiento) será limitar la segunda fase a un trabajo de campo con limitación temporal o a una cantidad fija de nuevos analizados por cada positivo encontrado en la primera.

Continuando con el ejemplo anterior, y una vez finalizada la Fase I, se pasa a la Fase II (estudio secuencial), en la que:

Etapa 3. Se pide a los positivos que hagan una lista de las personas empadronadas en Madrid con las que haya tenido contacto físico en los últimos 15 días.

Etapa 4. Se contacta a las nuevas personas para darles cita en los centros de salud y hacer las pruebas PCR y ELISA y los test rápidos.

Obviamente, el tamaño de muestra obtenido en las dos fases será superior a 250. Según la primera ronda de la encuesta ENE-COVID19, el porcentaje de españoles que habían pasado la enfermedad en mayo de 2020 se situa en torno al 5 % (11.3 % en el caso de Madrid). Esta estimación es coherente con el resultado publicado por [2], p. 6, el pasado 30 de marzo, en el que se daba un intervalo de credibilidad para la proporción de infectados que oscilaba entre el 3.7 % y el 41 %.

2.4. Implementación

Para poder implementar estos diseños, es necesaria la colaboración de oficinas de estadística, especialmente del INE, pudiendo participar oficinas municipales o autonómicas. Hay que usar información protegida por la actual Ley de Protección de Datos, pero que está disponible en las diferentes oficinas de estadística. En concreto, se necesitan los siguientes datos:

- El marco poblacional para poder extraer la muestra.
- Tamaños poblacionales (viviendas y personas) de los estratos y de los conglomerados para poder calcular las probabilidades de inclusión y calcular estimadores de medias y totales.

No obstante, aún sin la participación de ninguna oficina de estadística:

- Se podrá imitar el muestreo anterior usando la información cartográfica disponible.
- Como mucho, se podrán aplicar las fórmulas del muestreo aleatorio estratificado. En el caso de Madrid, los tamaños de los distritos están disponibles en la web estadística del ayuntamiento, pero no se dispone de los tamaños de las secciones censales.

2.5. Tamaño muestral

En este apartado se presentan unos cálculos sencillos para la determinación del tamaño muestral inicial de la encuesta. Se podrán hacer cálculos más específicos cuando se fije el diseño muestral final a implementar.

Sea una población U de tamaño N, donde se hace un muestreo aleatorio simple sin reemplazamiento de tamaño n. El intervalo de confianza (IC) a nivel

$(1-\alpha)100\,\%$ para una proporción poblacional $P \in \{0.1, \ldots, 0.5\}$, que usa la aproximación a la distribución normal (denotando el cuantil r de esta distribución por z_r), es

$$IC = \left(p - z_{1-\frac{\alpha}{2}} \sqrt{\frac{pq}{n-1}\frac{N-n}{N}} - \frac{1}{2n},\ p + z_{1-\frac{\alpha}{2}} \sqrt{\frac{pq}{n-1}\frac{N-n}{N}} + \frac{1}{2n} \right).$$

donde p es la proporción muestral y $q = 1 - p$. La semi-longitud del IC y el tamaño muestral sin corrección por continuidad son

$$d = p + z_{1-\frac{\alpha}{2}} \sqrt{\frac{pq}{n-1}\frac{N-n}{n}} + \frac{1}{2n}, \quad n_* = \frac{(d^2 + pqz^2_{1-\frac{\alpha}{2}})N}{d^2 N + pqz^2_{1-\frac{\alpha}{2}}}.$$

Suponiendo que $\frac{N-n}{N} \approx 1$, la Tabla 1 muestra las semi-longitudes d (filas 2 y 3, con $n = 250$ y $n = 1250$), y los tamaños muestrales n_* (filas 4 y 5, con d =0.02 y d =0.01), de los ICs para distintos valores de P y α =0.05.

Tabla 1: Semi-longitudes d y tamaños muestrales n_* de 95 %-ICs para P.

P	0.010	0.025	0.075	0.100	0.125	0.175	0.200
$n = 250$	0.0124	0.0194	0.0328	0.0374	0.0412	0.0473	0.0498
$n = 1250$	0.0055	0.0087	0.0146	0.0167	0.0184	0.0211	0.0222
$d = 0{,}02$	97	237	671	870	1057	1395	1546
$d = 0{,}01$	384	943	2680	3477	4225	5576	6179

Aunque los resultados de la Tabla 1 son específicos para un muestreo aleatorio simple sin reemplazamiento, son orientativos y de utilidad para cualquier otro diseño.

3. Diseño bietápico no secuencial

El contenido de este apartado está inspirado en el informe técnico de 2006 de la encuesta de población activa que realiza el INE para obtener información sobre el mercado laboral español. Adaptar un diseño del INE al estudio del COVID-19 facilitaría su rápida implementación por parte de una oficina de estadística. El informe técnico *Encuesta de Población Activa* [1] se puede descargar en: `https://www.ine.es/inebaseDYN/epa30308/docs/epa05_disenc.pdf`

Este apartado introduce un diseño bietápico no secuencial; es decir, un diseño que sólo aplica la Fase I. El muestreo se realiza en municipios (ciudades) que sean capital de provincia o que tengan un tamaño equivalente a la capital. El diseño divide el municipio en H estratos formados por distritos (barrios) vecinos. Cada estrato tiene un equipo propio para el trabajo de campo (contacto telefónico, toma de datos, realización de tests y digitalización).

Las unidades de primera etapa son las secciones censales. Para su selección se definen estratos dentro de los municipios. Dentro de cada estrato h, $h = 1, \ldots, H$, se seleccionan m_h secciones censales sin reemplazamiento y probabilidad proporcional al número de viviendas principales (según los datos del último Censo o Padrón). En el ejemplo de la sección anterior, se tiene $H = 5$ y $m_h = 5$.

Las unidades de segunda etapa son las viviendas familiares principales y alojamientos fijos. Dentro de cada sección seleccionada en primera etapa, se extrae un número fijo (por ejemplo, 10) de viviendas mediante la aplicación de un muestreo aleatorio simple.

Dentro de las unidades de segunda etapa no se realiza submuestreo alguno. Se aplican los tests serológicos, y se cumplimenta el correspondiente cuestionario, a todas las personas que tengan su residencia habitual en las viviendas seleccionadas.

Adoptamos la siguiente notación:

1. *Subíndices*: h para estratos, a para secciones, v para viviendas y j para individuos.
2. *Población y muestra*: U y s, con tamaños N y n respectivamente.
3. *Totales poblacionales*: V_{ha} y V_h son totales de viviendas familiares principales. El primero en la sección a del estrato h, y el segundo en el estrato h.
4. *Totales muestrales*: m_h es el número de secciones censales seleccionadas en el estrato h.

Los cálculos siguientes se efectúan bajo el supuesto de que se extrae un número fijo de 10 viviendas con reemplazamiento. La probabilidad de selección de la vivienda v de la sección a del estrato h es

$$P(v \in s_{ha}) = P(a \in s_h)P(v \in s_{ha} \,/\, a \in s_h) \approx m_h \frac{V_{ha}}{V_h} \frac{10}{V_{ha}} = \frac{10\, m_h}{V_h}.$$

Puesto que todos los individuos de una vivienda seleccionada son encuestados, la probabilidad de inclusión de un individuo j coincide con la de su vivienda. Así pues, la probabilidad de selección del individuo j de la vivienda v del estrato h es

$$\pi_j = \frac{10\, m_h}{V_h} \triangleq \pi_h. \tag{3.1}$$

Con lo cual, las muestras son auto-ponderadas dentro de los estratos. Las probabilidades π_h se invierten para dar lugar a los pesos teóricos del diseño muestral; es decir,

$$w_j = \frac{1}{\pi_j} = \frac{V_h}{10\, m_h} = w_h, \quad j \in s_h.$$

La variable y de interés, medida en el individuo j, es

$$y_j = \begin{cases} 1 & \text{si } j \text{ da positivo en test ELISA de anticuerpos anti SARS-COV-2,} \\ 0 & \text{en caso contrario.} \end{cases} \tag{3.2}$$

Para cada municipio U de tamaño N, los parámetros de interés son el total de positivos y la correspondiente proporción; es decir,

$$Y = \sum_{j \in U} y_j, \quad P = \overline{Y} = \frac{Y}{N}.$$

La variable dicotómica relacionada, z_j, registra el resultado del test rápido de anticuerpos anti SARS-COV-2.

3.1. Estimador no calibrado

El estimador no calibrado del total Y de la variable y en el municipio U es

$$\hat{Y} = \sum_{h \in U} \frac{N_h}{\widehat{N}_h} \sum_{v \in s_h} \sum_{j \in v} w_j y_j,$$

donde

- N_h es el número de individuos que residen en viviendas familiares en el estrato h. El valor de N_h se toma de la proyección de población en el estrato h referida a la mitad del trimestre en la que se extrae la muestra.
- $\widehat{N}_h = \sum_{v \in s_h} \sum_{j \in v} w_j = w_h \sum_{v \in s_h} N_v = w_h n_h,$
- N_v es el número de individuos que residen en la vivienda v.
- $n_h = \sum_{v \in s_h} N_v$ es el número de individuos de la muestra en el estrato h.

El estimador no calibrado es un estimador de razón que admite la expresión alternativa

$$\hat{Y} = \sum_{h \in U} N_h \left(\frac{1}{\widehat{N}_h} \sum_{j \in s_h} w_j y_j \right) \triangleq \sum_{h \in U} N_h \widehat{\overline{Y}}_h;$$

es decir, es un estimador post-estratificado donde los grupos de post-estratificación son los estratos del diseño. Los estimadores $\hat{Y}$ y $\widehat{\overline{Y}}$ se pueden escribir de la siguiente forma

$$\hat{Y} = \sum_{h \in U} \sum_{j \in s_h} \frac{N_h\, w_h}{\widehat{N}_h} y_j = \sum_{j \in s} w_j^b y_j, \quad \widehat{\overline{Y}} = \frac{\hat{Y}}{\hat{N}}, \quad \hat{N} = \sum_{j \in s} w_j^b,$$

donde

$$w_j^b = w_j^b(s) = \frac{N_h\, w_h}{\widehat{N}_h} = \frac{N_h}{n_h}, \qquad \text{si} \quad j \in s_h.$$

Conviene recalcar el hecho de que los pesos no calibrados w_j^b dependen de la muestras, puesto que $n_h = n_h(s)$. Es decir, los pesos no calibrados y los tamaños muestrales son aleatorios. En cambio, los pesos teóricos w_j son valores fijos.

3.2. Calibración de pesos

Consideremos el estimador no calibrado

$$\hat{Y} = \sum_{j \in s} w_j^b y_j.$$

Supongamos que se dispone de K variables objetivo cuyos totales

$$X_k = \sum_{j \in P} x_{jk}, \quad k = 1, \ldots, K,$$

son conocidos para la población. El problema de *calibración* consiste en encontrar un nuevo estimador

$$\hat{Y}_c = \sum_{j \in s} w_j^c\, y_j,$$

donde los pesos calibrados w_j^c cumplan las siguientes condiciones:

1. Sean próximos a los pesos iniciales w_j^b;
2. Verifiquen las ecuaciones de equilibrado

$$\sum_{j \in s} w_j^c\, x_{jk} = X_k, \quad k = 1, \ldots, K.$$

El planteamiento del problema es encontrar unos valores w_j^c que minimicen la expresión

$$\sum_{j \in s} w_j^b\, \phi(w_j^c / w_j^b), \qquad \text{sujeto a} \quad \sum_{j \in s} w_j^c\, x_{jk} = X_k, \;\; k = 1, \ldots, K, \tag{3.3}$$

siendo ϕ una función monótona decreciente a la izquierda de $x = 1$, monótona creciente a la derecha de $x = 1$, y tal que $\phi(1) = 1$. Funciones de uso frecuente son $\phi(x) = (x-1)^2/2$ o $\phi(x) = x \log x$, $x \in R$.

Observación. El INE usa la función de distancia lineal truncada para evitar las soluciones negativas. Como variables auxiliares para la EPA, utiliza:

- Población de 16 o más años por grupos de edad y sexo a nivel de Comunidad Autónoma.

- Población de 16 y más años por provincia.

En el muestreo en municipios, se puede calibrar a

- Población por grupos de edad y sexo en el municipio.

Los estimadores calibrados estiman correctamente la población por grupo y edad. Para la solución práctica del problema de calibración, se puede utilizar la función calib del paquete sampling de R: `https://rdrr.io/cran/sampling/`

Alternativamente, la calibración se puede hacer usando el software CALMAR que se puede bajar de la dirección web `http://www.insee.fr/fr/nom_def_met/outils_stat/calmar/cal_res.htm`

3.3. Estimador calibrado

El estimador calibrado del total y de la media de la variable y, en el municipio estudiado, es

$$\hat{Y}^c = \sum_{j \in s} w_j^c y_j, \quad \hat{\overline{Y}}^c = \frac{\hat{Y}^c}{\hat{N}^c}, \quad \hat{N}^c = \sum_{j \in s} w_j^c,$$

donde los w_j^c son los pesos calibrados.

La varianza de $\hat{Y}^c$ se puede estimar por procedimientos de remuestreo (semi-muestras reiteradas, Jackknife o bootstrap). El informe técnico de la EPA explica como extraer las semi-muestras. Herrador [3] adapta los remuestreos bootstrap y Jackknife al diseño muestral de la EPA. Conviene recalcar el hecho de que cada remuestra debe ser calibrada para aplicar el método de remuestreo correctamente.

Alternativamente, se pueden usar los estimadores

$$\hat{V}_\pi(\hat{Y}^c) = \sum_{j \in S} w_j(w_j - 1)\big(y_j - \hat{\overline{Y}}^c\big)^2, \quad \hat{V}_\pi(\hat{\overline{Y}}^c) = \hat{V}\big(\hat{Y}^c\big)/\big(\hat{N}^c\big)^2. \tag{3.4}$$

Estas fórmulas se obtienen de [5], pp. 43, 185 y 391, con las simplificaciones $w_j = 1/\pi_j$, $\pi_{jj} = \pi_j$ y $\pi_{ij} = \pi_i \pi_j$, $i \neq j$, en las probabilidades de inclusión de primer (π_j) y segundo (π_{ij}) orden.

3.4. Imputación de valores ELISA

Solamente un fracción de los participantes en la muestra aceptarán que se les haga una venopunción. Por tanto, la variable y, definida en (3.2), presentará datos faltantes. Sin embargo, la variable dicotómica z que contiene el resultado del test rápido estará registrada para todas las personas seleccionadas en el estudio. Una forma sencilla de mejorar la estimación de la media poblacional de la variable y por provincias y grupos de sexo-edad es imputar los valores faltantes mediante el modelo logístico mixto empleado por [4] para estimar proporciones de pobreza en áreas pequeñas.

Sean los grupos de edad < 10, 10-19, 20-34, 35-49, 50-64 y > 65 años. Para cada provincia d (incluyendo Ceuta y Melilla), grupo de sexo-edad t y participante j, suponemos que

$$y_{dtj}|_{v_{1,d},v_{2,dt}} \sim \text{Bin}(1, p_{dtj}), \quad d = 1, \ldots, D,\ t = 1, \ldots, T,\ j = 1, \ldots, n_{dt}, \quad (3.5)$$

siendo $D = 52$, $T = 12$ y n_{dt} el tamaño muestral del cruce provincia-grupo. Para el parámetro natural asumimos el nexo logístico

$$\eta_{dtj} = \log \frac{p_{dtj}}{1 - p_{dtj}} = \mathbf{x}_{dtj}\boldsymbol{\beta} + \phi_1 v_{1,d} + \phi_{2t} v_{2,dt}, \quad (3.6)$$

donde $\phi_1 > 0$ y $\phi_{2t} > 0$, $t = 1, \ldots, T$, son parámetros de desviación típica, $\boldsymbol{\beta} = \underset{1 \le k \le p}{\text{col}} (\beta_k)$ es el vector de parámetros de regresión y $\mathbf{x}_{dtj} = \underset{1 \le k \le p}{\text{col}'} (x_{dtjk})$ es el vector de variables explicativas, con valores disponibles en todos los individuos de la muestra. Finalmente, los efectos aleatorios, $v_{1,d}$, $v_{2,dt}$, $d = 1, \ldots, D$, $t = 1, \ldots, T$, son i.i.d. $N(0, 1)$.

Se propone usar el resultado z del test rápido como variable explicativa, imputar los valores faltantes de y usando las predicciones $\hat{p}_{dtj}$ del modelo ajustado y posteriormente aplicar los estimadores de esta sección al vector completado de valores de la variable y.

4. Diseño bietápico secuencial en la población

Thompson [6] introduce un muestreo por conglomerados secuencial multifásico. En esta sección describimos una posible adaptación de ese diseño muestral al problema de estimar la proporción de la población portadora de anticuerpos en un municipio dado, considerando la Fase I y Fase II explicadas en la Sección 2.

La Fase I conlleva la extracción de una muestra s_1 de tamaño n_1 con el diseño muestral descrito en la Sección 3. Una vez realizados los tests y las entrevistas a las personas de la muestra s_1, se pasa a la siguiente fase.

La Fase II fija su atención en las personas que viven en las viviendas seleccionadas en s_1 y que han dado positivo en el test PCR (COVID-19-positivas). Se entrevista a aquellas personas que no hayan sido previamente contactadas y que pertenecen a la 1-red de contacto físico de las personas COVID-19-positivas de s_1. De este modo se obtiene una segunda muestra que denotamos s_2 con intersección vacía con s_1.

La 1-red de contacto físico de una persona j es ella misma, si la persona ha dado negativo. En cambio, si ha dado positivo, su 1-red coincide con la 1-red de la vivienda v en la que habita. La 1-red de contacto físico de una vivienda v está formada por el conjunto de habitantes del municipio que han tenido un contacto estrecho con alguno de sus miembros. El contacto estrecho se refiere

a un distanciamiento inferior a dos metros, sin protección, durante un tiempo mínimo de 15 minutos y en los 15 días anteriores a la toma de datos. Se incluye en esa 1-red a todas las personas que comparten una misma vivienda. Por tanto, si las personas de la vivienda v_1 están en la 1-red de la vivienda v_2, entonces las personas de la vivienda v_2 están en la 1-red de la vivienda v_1.

En este diseño tenemos una Fase III que consiste en hacer los tests COVID-19 y entrevistar a aquellas personas que pertenecen a la 1-red de contacto físico de las personas COVID-19-positivas que viven en las viviendas seleccionadas en s_2 y que no hayan sido previamente contactadas. De este modo se obtiene la muestra s_3. Se verifica que s_1, s_2 y s_3 son disjuntas dos a dos.

Se itera el procedimiento descrito hasta que en una etapa dada, los candidatos a entrar en la muestra s_{K+1} sean todos personas que ya pertenecen a algunas de las muestras anteriores $s_1, \ldots, s_K$, o bien, todos los entrevistados dan negativo. Por tal motivo $s_{K+1} = \emptyset$ y la muestra final es $s = \cup_{k=1}^{K} s_k$

El cluster C_i (diferente del conglomerado *sección censal*), asociado a la persona $i \in s_1$, es el subconjunto de la población que entraría en la muestra final s por conexiones 1-red con $i \in s_1$, incluyendo las personas de la propia vivienda. Si no hay ningún COVID-19-positivo en una vivienda dada, entonces el cluster asociado a cualquiera de sus miembros es el conjunto de personas de la vivienda.

La red de contacto $R_i \subset C_i$ es el subconjunto de la población que contiene a la persona $i \in s_1$ y que tiene la propiedad de que todos sus miembros estarían en la muestra final s, si cualquier de ellos estuviera en la muestra s_1 de la fase 1. Sea $m_i = |R_i|$ el número de personas en R_i. Sea $\bar{y}_i^* = \frac{1}{m_i} \sum_{j \in R_i} y_j$.

Suponiendo que s_1 se extrae con muestreo aleatorio simple sin reemplazamiento, en lugar de con el diseño muestral de la Sección 3, Thompson [6] propone estimar $\overline{Y}$ con la modificación del estimador Hansen-Hurwitz

$$\hat{\overline{Y}}_T = \frac{1}{n_1} \sum_{i=1}^{n_1} \bar{y}_i^*.$$

Además, Thompson [6] asume que las redes de contacto R_i y sus tamaños m_i son conocidos para todos los individuos i de la muestra final s.

4.1. Metodología simplificada

Si la muestra s_1 se extrae con el diseño muestral descrito en la Sección 3, entonces la metodología de Thompson [6] es dificilmente aplicable al problema que nos ocupa. Hay dos elementos que no están resueltos: (1) Proponer el estimador para el diseño muestral de la Sección 3 y estudiar sus propiedades, y (2) Determinar R_i y m_i para toda unidad de la muestra s_1. El primer problema, require un estudio teórico no trivial. El segundo problema afecta a la aplicabilidad de la metodología.

Una manera rápida y sencilla de convertir la metodología de la Sección 3

en *secuencial aplicable* es simplificar la propuesta de Thompson [6] añadiendo hipótesis más fuertes y limitando el diseño muestral a 2 etapas. El adjetivo aplicable se usa en el sentido de poder cumplir con los requerimientos presupuestarios y temporales exigidos en el trabajo de campo de la encuesta.

Si paramos en la etapa 2 y no construimos R_i, entonces no se puede calcular $\hat{\overline{Y}}_T$. En ese caso la muestra final es $s_{12} = s_1 \cup s_2$. Las probabilidades de inclusión de las unidades $j \in s_1$ son las mismas que en (3.1); es decir

$$\pi_j = \frac{10\, m_h}{V_h} \triangleq \pi_{1h}, \quad j \in s_1. \tag{4.1}$$

Aproximamos la probabilidad de inclusión de las unidades $j \in s_2$. Sean v_1 y v_2 dos viviendas del estrato h. Suponemos que: (1) $j \in v_2 \subset s_2$, (2) j unicamente está en la 1-red contacto físico de las personas $i \in v_1 \subset s_1$. Entonces, damos la siguiente aproximación

$$\begin{aligned} \pi_j &= P(v_2 \in s_{ha12}) \approx P(v_2 \in s_{ha1}) + P(v_2 \notin s_{ha1}) P(v_1 \in s_{ha1}) \\ &= \frac{10\, m_h}{V_h} + \Big(1 - \frac{10\, m_h}{V_h}\Big) \frac{10\, m_h}{V_h} \triangleq \pi_{2h}, \quad j \in s_2. \end{aligned} \tag{4.2}$$

Observamos que las muestras no son auto-ponderadas dentro de los estratos. Las probabilidades π_{kh}, $k = 1, 2$, se invierten para dar lugar a los pesos teóricos del diseño muestral; es decir,

$$w_j = \frac{1}{\pi_j}, \quad j \in s_h.$$

Las aproximaciones (4.1) y (4.2) dan mayor peso muestral a las unidades seleccionadas en la muestra s_1. De ese modo, al calcular la estimación de la prevalencia como media ponderada, tales unidades son más determinantes para el resultado final. Ello disminuye el posible sesgo positivo.

A partir de este punto, son válidos los desarrollos de la Sección 3. En este caso, la calibración de los pesos muestrales es muy aconsejable para disminuir el sesgo producido por la aproximación (4.2). Por otra parte, la calibración siempre es recomendable pues reduce la varianza de los estimadores no calibrados.

5. Diseño bietápico secuencial en los conglomerados

En esta sección se estudia otro caso particular distinto del anterior, donde la cadena de contagios se restringe sólo a los conglomerados. Ello permite utilizar la fórmulas de Thompson [6] para el muestreo por conglomerados secuencial multi-fásico. Describiremos una posible adaptación de ese diseño muestral al problema de estimar la proporción de la población portadora del virus (o de anticuerpos) en un municipio dado.

La Fase I conlleva la extracción de una muestra $s_1 = \cup_{h=1}^{H} \cup_{a=1}^{m_h} s_{ha1}$ con el diseño muestral descrito en la Sección 3. La submuestra correspondiente a la sección censal a del estrato h es s_{ha1}. Los correspondientes tamaños muestrales son n_1 y n_{1ah}. Una vez realizados los tests y las entrevistas a todas las personas de las muestras s_{ha1}, $h = 1, \ldots, H$, $a = 1, \ldots, m_h$, se pasa a la siguiente fase.

La Fase II fija su atención en las personas que viven en las viviendas seleccionadas en cada s_{ha1} y que han dado positivo en el test PCR (COVID-19-positivas). Se entrevista a aquellas personas que no hayan sido previamente contactadas y que pertenecen a la 1-red de contacto físico de las personas COVID-19-positivas de s_{ha1}. De este modo se obtiene una segunda muestra que denotamos s_{ha2} con intersección vacía con s_{ha1}.

La 1-red de contacto físico de una persona i es ella misma, si la persona ha dado negativo. En cambio, si ha dado positivo, su 1-red coincide con la 1-red de la vivienda v en la que habita. La 1-red de contacto físico de una vivienda v está formada por el conjunto de habitantes del conglomerado (sección censal) que han estado sin protección durante más de 15 minutos, en los 15 días anteriores a la toma de datos, a menos de dos metros de distancia de alguno de sus miembros. Se incluye en esa 1-red a todas las personas que comparten una misma vivienda. Por tanto, si las personas de la vivienda v_1 están en la 1-red de la vivienda v_2, entonces las personas de la vivienda v_2 están en la 1-red de la vivienda v_1.

Para cada h y a, la Fase III consiste en hacer los test COVID-19 y entrevistar a aquellas personas que pertenecen a la 1-red de contacto físico de las personas COVID-19-positivas que viven en las viviendas seleccionadas en s_{ha2} y que no hayan sido previamente contactadas. De este modo se obtiene la muestra s_{ha3}. Se verifica que s_{ha1}, s_{ha2} y s_{ha3} son disjuntas dos a dos.

Se itera el procedimiento descrito hasta que en una etapa dada, los candidatos a entrar en la muestra $s_{ha(K+1)}$ sean todos personas que ya pertenecen a algunas de las muestras anteriores $s_1, \ldots, s_{haK}$, o bien, todos los entrevistados dan negativo. Por tal motivo $s_{ha(K+1)} = \emptyset$ y la muestra final del conglomerado a del estrato h es $s_{ha} = \cup_{k=1}^{K} s_{hak}$, $h = 1, \ldots, H$, $a = 1, \ldots, m_h$. La muestra global es $s = \cup_{h=1}^{H} \cup_{a=1}^{m_h} s_{ha}$.

El cluster C_{hai} (diferente del conglomerado *sección censal*), asociado a la persona $i \in s_{ha1}$, es el subconjunto del conglomerado (al que pertence i) que entraría en la muestra final s por conexiones 1-red con $i \in s_{ha1}$, incluyendo las personas de la propia vivienda. Si no hay ningún COVID-19-positivo en una vivienda dada, entonces el cluster asociado a cualquiera de sus miembros es el conjunto de personas de la vivienda.

La red de contacto $R_{hai} \subset C_{hai}$ es el subconjunto de personas de C_{hai} que tiene la propiedad de que todos sus miembros estarían en la muestra final s_{ha}, si cualquier de ellos estuviera en la muestra s_{ha1} de la fase 1. Sea $r_{hai} = |R_{hai}|$

el número de personas en R_{hai}. Sea

$$\bar{y}^*_{hai} = \frac{1}{r_{hai}} \sum_{j \in R_{hai}} y_j.$$

Suponiendo que s_{ha1} se extrae con muestreo aleatorio simple sin reemplazamiento, Thompson [6] propone estimar $\overline{Y}_{ha}$, y la correspondiente varianza, con la modificación del estimador Hansen-Hurwitz

$$\hat{\overline{Y}}_{ha} = \frac{1}{n_{ha1}} \sum_{i=1}^{n_{ha1}} \bar{y}^*_{hai}, \quad \hat{V}_\pi\big(\hat{\overline{Y}}_{ha}\big) = \frac{1}{n_{ha1}(n_{ha1}-1)} \sum_{i=1}^{n_{ha1}} \big(\bar{y}^*_{hai} - \hat{\overline{Y}}_{ha}\big)^2.$$

Para el total, se tiene

$$\hat{Y}_{ha} = \sum_{i=1}^{n_{ha1}} \bar{y}^*_{hai}, \quad \hat{V}_\pi\big(\hat{Y}_{ha}\big) = \frac{n_{ha1}}{(n_{ha1}-1)} \sum_{i=1}^{n_{ha1}} \big(\bar{y}^*_{hai} - \hat{\overline{Y}}_{ha}\big)^2.$$

Además, Thompson [6] asume que las redes de contacto R_{hai} y sus tamaños r_{hai} son conocidos para todos los individuos i de la muestra final s_{ha}, $h = 1, \ldots, H$, $a = 1, \ldots, m_h$.

Si el muestreo se realiza con reemplazamiento, se tiene la aproximación.

$$\begin{aligned} P(a \in s_h) &= 1 - P(a \notin s_h) = 1 - \left(1 - \frac{V_{ha}}{V_a}\right)^{m_h} \\ &= 1 - \left\{ \sum_{k=0}^{m_h} \binom{m_h}{k} (-1)^k \left(\frac{V_{ha}}{V_a}\right)^k \right\} \approx m_h \frac{V_{ha}}{V_h}. \end{aligned}$$

Por tanto, un estimador del total Y_h es

$$\hat{Y}_h = \sum_{a=1}^{m_h} \frac{V_h}{m_h V_{ha}} \hat{Y}_{ha},$$

y un estimador del total Y es

$$\begin{aligned} \hat{Y} &= \sum_{h=1}^{H} \frac{N_h}{\widehat{N}_h} \hat{Y}_h = \sum_{h=1}^{H} \frac{N_h}{\widehat{N}_h} \sum_{a=1}^{m_h} \frac{V_h}{m_h V_{ha}} \hat{Y}_{ha} \\ &= \sum_{h=1}^{H} \frac{N_h}{\widehat{N}_h} \sum_{a=1}^{m_h} \frac{V_h}{m_h V_{ha}} \frac{1}{n_{ha1}} \sum_{i=1}^{n_{ha1}} \frac{1}{r_{hai}} \sum_{j \in R_{hai}} y_j. \end{aligned}$$

Si usamos la notación y_{haij} para el valor de la variable y medida en el individuo final j de la red de contacto R_{hai} de la persona $i \in s_{ha1}$, $h = 1, \ldots, H$, $a =$

$1, \ldots, m_h$, entonces un estimador no calibrado de Y es

$$\hat{Y} = \sum_{h=1}^{H} \sum_{a=1}^{m_h} \sum_{i=1}^{n_{ha1}} \sum_{j \in R_{hai}} \frac{N_h V_h}{\widehat{N}_h m_h V_{ha} n_{ha1} r_{hai}} y_{haij} \triangleq \sum_{h=1}^{H} \sum_{a=1}^{m_h} \sum_{i=1}^{n_{ha1}} \sum_{j \in R_{hai}} w^b_{haij} y_{haij}.$$

El correspondiente estimador de $\overline{Y}$ es

$$\hat{\overline{Y}} = \frac{\hat{Y}}{\hat{N}}, \quad \hat{N} = \sum_{h=1}^{H} \sum_{a=1}^{m_h} \sum_{i=1}^{n_{ha1}} \sum_{j \in R_{hai}} w^b_{haij}.$$

El estimador calibrado del total y de la media de la variable y, en el municipio estudiado, es

$$\hat{Y}^c = \sum_{h=1}^{H} \sum_{a=1}^{m_h} \sum_{i=1}^{n_{ha1}} \sum_{j \in R_{hai}} w^c_{haij} y_{haij}, \ \hat{\overline{Y}}^c = \frac{\hat{Y}^c}{\hat{N}^c}, \ \hat{N}^c = \sum_{h=1}^{H} \sum_{a=1}^{m_h} \sum_{i=1}^{n_{ha1}} \sum_{j \in R_{hai}} w^c_{haij},$$

donde los w^c_{haij} son los pesos calibrados a los totales poblacionales del municipio por grupos de edad y sexo.

Podemos estimar las varianzas de $\hat{Y}^c$ y $\hat{\overline{Y}}^c$ por cualquiera de los procedimientos descritos en la Sección 3.3.

Agradecimientos

Queremos agradecer al Prof. Javier Llorca Díaz, Universidad de Cantabria, por los comentarios que hizo a una versión preliminar de esta propuesta, aportando ideas que la mejoraron considerablemente. No obstante, todos los errores que pueda contener han de achacarse a los autores de la misma.

Referencias

[1] Encuesta de Población Activa (2006). Informe técnico. Área de diseño de muestras y evaluación de resultados. Instituto Nacional de Estadística.

[2] Flaxman, S., Mishra, S. Gandy, A. et al. (2020). Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Imperial College London.

[3] Herrador, M. Morales, D., Esteban, M.D., Sánchez, A., Santamaría, L., Marhuenda, Y., Pérez, A. (2008). Sampling design variance estimation of small area estimators in the Spanish Labour Force survey. *SORT*, 32, 2, 177-198.

[4] Hobza, T., Morales, D., Santamaría, L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST*, 27 270–294.

[5] Särndal, C.E., Swensson, B., Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag.

[6] Thompson, S.K. (1990). Adaptive Cluster Sampling. *Journal of the American Statistical Association*, 85, 412, 1050-1059.

[7] Thompson, S.K. (1991). Stratified Adaptive Cluster Sampling *Biometrika*, 78, 2, 389-397.

Acerca de los autores

Domingo Morales González es Catedrático de Universidad en el Departamento de Estadística, Matemáticas e Informática de la Universidad Miguel Hernández de Elche, así como investigador del Instituto Universitario Centro de Investigación Operativa. Su principal área de investigación es la estimación en áreas pequeñas y la modelización estadística, con particular interés en desarrollos metodológicos para la estadística pública. Ha sido Secretario General de la SEIO desde octubre de 1994 hasta noviembre de 2001, presidente de la SEIO desde octubre de 2004 hasta septiembre de 2007 y co-editor jefe de la revista TEST durante los años 2009 a 2013. Actualmente es editor asociado de las revistas TEST y Computational Statistics and Data Analysis.

María José Lombardía Cortiña es titular de Universidad en el Departamento de Matemáticas de la Universidade da Coruña (UDC), investigadora del Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC) y pertenece al grupo de investigación Modelización, Optimización e Inferencia Estadística (MODES). Sus principales áreas de investigación son la estimación en áreas pequeñas, modelos mixtos, métodos de remuestreo bootstrap e inferencia no paramétrica y la modelización estadística, con particular interés en desarrollos metodológicos para la estadística pública. Fue secretaria y miembro del Consejo Ejecutivo de la Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións (SGAPEIO), y miembro del Consejo Ejecutivo de la Sociedad Española de Biometría (SEB). Además, fue miembro del Comité de Dirección del CITIC y durante más de 7 años estuvo al frente de la Unidad de Análisis y Gestión de datos de la UDC. Actualmente es Vicerrectora de Estudiantes, Planificación y Empleabilidad.

Ricardo Fraiman es profesor titular del Centro de Matemática de la Universidad de la República, Uruguay, y miembro de la Academia Nacional de Ciencias

del Uruguay. Antes se desempeñó en el Departamento de Matemática de la facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires, y en la Universidad de San Andrés. Su área de investigación se ha centrado en estimación no paramétrica, en la estadística robusta, estadística de datos funcionales, aprendizaje supervisado y no supervisado, estimación de conjuntos y funcionales asociados, estimación en redes y aplicaciones en neurociencias, y la aplicación de proyecciones aleatorias en problemas de alta dimensión. Fue editor asociado de TEST y Bernoulli, y es editor asociado de ALEA. Ha colaborado de forma sistemática con varios colegas españoles en trabajos de investigación conjunta.

Juan Antonio Cuesta Albertos es Catedrático de Universidad en el Departamento de Matemáticas, Estadística y Computación de la Universidad de Cantabria. Su investigación se centra en la estadística robusta, análisis de datos funcionales, problema del transporte y la aplicación de proyecciones aleatorias en problemas de alta dimensión. Fue co-editor jefe de la revista TEST durante los años 2002 a 2004. Actualmente es editor asociado del Journal of the American Statistical Association. Dedica parte de su actividad a la impartición de charlas divulgativas.