UNIVERSIDAD DE GRANADA

PROGRAMA DE DOCTORADO EN ESTADÍSTICA MATEMÁTICA Y APLICADA

Tesis Doctoral

Aportaciones en encuestas no probabilísticas y encuestas web

Héctor Salomón Mullo Guaminga

Director de Tesis:

Prof. Ismael Ramón Sánchez Borrego

Tutora de Tesis:

Prof. María del Mar Rueda García

Granada 2021

Editor: Universidad de Granada. Tesis Doctorales
Autor: Héctor Salomón Mullo Guaminga
ISBN: 978-84-1117-076-5
URI: http://hdl.handle.net/10481/71173

Dedicado a mi querida familia y en especial a Sarahí la princesa de mi vida.

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

En el devenir de nuestra existencia es importante valorar y agradecer la ayuda brindada por los que nos rodean en la consecución de nuestros sueños. Y en especial sobre esta tesis doctoral que ha sido el sueño más grande que eventualmente supero mis miedos. Tengo el privilegio de reconocer a todas las personas que ayudaron a hacer posible la culminación de este logro.

En primer lugar, quiero agradecer a mis directores de tesis D. Ismael Ramón Sánchez Borrego y Dña. María del Mar Rueda García quienes con sus conocimientos y experiencia han guiado todo el transcurrir de esta investigación. Gracias por su invaluable ayuda, comprensión y calidad humana.

También quiero dar las gracias a la Confederación Nacional de Organizaciones Campesinas Indígenas y Negras del Ecuador en cabeza de D. Santos Villamar y D. Miguel Quijijen quienes me brindaron su apoyo para la encuesta de minorías étnicas.

Finalmente, agradezco afectuosamente a mi familia por estar junto a mi apoyándome incondicionalmente en especial a mi esposa e hija junto con mis padres y hermanos.

Granada, junio de 2021.

Resumen

Antecedentes: El propósito principal del muestreo estadístico es obtener conocimiento sobre una población utilizando un subconjunto pequeño y accesible de individuos seleccionados. Este objetivo generalmente se aborda al elegir una muestra representativa, utilizando la probabilidad de selección de cada individuo determinada por una lista completa de la población objetivo (marco muestral). Sin embargo, para muchas poblaciones ocultas o de difícil acceso importantes para la Salud pública y la Sociología, como las que están en riesgo de contraer VIH, la comunidad LGBTI, minorías étnicas, inmigrantes y niños de la calle, etc, la probabilidad de selección de los individuos no se puede determinar, debido a que no existe un marco muestral o su construcción es extremadamente costosa y poco práctica debido a su naturaleza. El muestreo dirigido por los participantes (RDS) se introdujo por Heckathorn [4] y es uno de los métodos más utilizados cuando se toman muestras de poblaciones ocultas o de difícil acceso. La metodología RDS combina un esquema de muestreo de bola de nieve mejorado, con un modelo matemático que puede producir estimaciones no sesgadas de la población, cuando se cumplen algunas suposiciones sobre el proceso de reclutamiento. En RDS existen estrategias bien desarrolladas para estimar medias y prevalencias. Además, dentro del desarrollo de métodos de regresión, existen trabajos muy interesantes basados en la modelización que intentan considerar la agrupación de redes subyacentes a la muestra y la homofilia inherente al proceso de reclutamiento. Sin embargo, no se tiene un método estandarizado para el modelado de regresión utilizando datos recopilados a través de RDS. También, en la literatura de trabajos que aplican la metodología RDS apenas hay trabajos realizados para muestrear poblaciones de minorías étnicas.

Objetivos: El objetivo de esta tesis es doble. Primero, poner en práctica una encuesta no probabilística mediante la metodología RDS, para recoger información sobre poblaciones indígenas y otras minorías étnicas de relevancia en Sudamérica. En segundo lugar, formular expresiones de estimadores de

covarianza y coeficientes de correlación en el contexto del muestreo de poblaciones difíciles de alcanzar.

Contribuciones: Se ha demostrado que, RDS es un método eficaz para analizar la estructura de las redes sociales de minorías étnicas conectadas en la web y se ha implementado con éxito como método de muestreo en la web en Ecuador, un país en el que las poblaciones minoritarias están muy estigmatizadas y subrepresentadas en las encuestas oficiales. Los métodos estadísticos se desarrollan en el contexto de regresión para RDS, en este sentido se ha propuesto un nuevo método de estimación de los pesos de la muestra para datos continuos. El enfoque de nuestro trabajo ha sido proponer un método para estimar parámetros no lineales con nuevos pesos muestrales. Se han derivado expresiones de las varianzas y también se ha demostrado que los estimadores propuestos tienen propiedades deseables.

Conclusiones: El muestreo dirigido por los participantes es una herramienta eficaz para muestrear en la web minorías étnicas urbanas en Ecuador. Sin embargo, debido a varias violaciones de los supuestos de RDS la interpretación de las estimaciones de población de la muestra de RDS en la web debe estar condicionado a estas incertidumbres. Los resultados sobre la dependencia entre variables continuas presentados en esta tesis, se suman a la creciente literatura sobre muestreo dirigido por los participantes, lo que permite a los investigadores obtener mejor información sobre las poblaciones ocultas de interés. Estos hallazgos son importantes para los métodos de regresión en el contexto de RDS y para muestrear minorías étnicas sin la limitación de las implementaciones físicas basadas en entrevistas.

Summary

Background: The main purpose of statistical sampling is to obtain knowledge about a population using a small and accessible subset of selected individuals. This objective is generally addressed by choosing a representative sample, using the probability of selection of each individual determined by a complete list of the target population (sampling frame). However, for many hidden or hard-to-reach populations important to Public Health and Sociology, such as those at risk of contracting HIV, the LGBTI community, ethnic minorities, immigrants and street children, etc., the probability of selection of individuals cannot be determined, because there is no sampling frame or its construction is extremely expensive and impractical due to its nature. Respondent-Driven Sampling (RDS) was introduced by Heckathorn [4] and is one of the most widely used methods when sampling from hidden or hard-to-reach populations. The RDS methodology combines an improved snowball sampling scheme with a mathematical model that can produce unbiased estimates of the population, when some assumptions about the recruitment process are met. In RDS there are well-developed strategies for estimating means and prevalences. In addition, within the development of regression methods, there are very interesting works based on modeling that attempt to consider the grouping of networks underlying the sample and the homophily inherent in the recruitment process. However, there is no standardized method for regression modeling using data collected through RDS. Also, in the literature of works that apply the RDS methodology there are hardly any works carried out to sample ethnic minority populations.

Objectives: The objective of this thesis is twofold. First, to implement a non-probabilistic survey using the RDS methodology to collect information on indigenous populations and other relevant ethnic minorities in South America. Second, formulate expressions of covariance estimators and correlation coefficients in the context of sampling hard-to-reach populations.

Contributions: RDS has been shown to be an effective method for analyzing the structure of ethnic minority social networks connected on the web and has been successfully implemented as a web sampling method in Ecuador, a country where minority populations are present. highly stigmatized and underrepresented in official surveys. The statistical methods are developed in the regression context for RDS, in this sense a new method of estimating the sample weights for continuous data has been proposed. The focus of our work has been to propose a method to estimate non-linear parameters with new sample weights. Expressions of the variances have been derived and the proposed estimators have also been shown to have desirable properties.

Conclusions: Respondent-Driven Sampling is an effective tool for sampling urban ethnic minorities in Ecuador on the web. However, due to various violations of the RDS assumptions the interpretation of the population estimates of the RDS sample on the web must be conditioned by these uncertainties. The results on the dependence between continuous variables presented in this thesis add to the growing literature on Respondent-Driven Sampling, allowing researchers to obtain better information on the hidden populations of interest. These findings are important for regression methods in the context of RDS and for sampling ethnic minorities without the limitation of physical implementations based on interviews.

Capítulo 1

Introducción

En muchas áreas de investigación, la obtención de una muestra representativa sobre el comportamiento y composición de un determinado grupo objeto de estudio, es un problema resuelto mediante métodos de muestreo tradicionales. Sin embargo, en algunas poblaciones con características muy específicas o de difícil acceso, tales técnicas no pueden ser aplicadas puesto que no es posible o es difícil construir un marco muestral que permita la selección de individuos mediante una probabilidad de selección. Una población difícil de alcanzar u oculta se define como un pequeño subgrupo de una población de interés para el invetigador, donde la población es dificil de acceder y además la población no tiene límites definidos, de tal manera que su tamaño exacto no puede ser conocido [1, 2].

A menudo, estas poblaciones son costosas de identificar y sus miembros suelen recelar de revelar su pertenencia a la misma, bien porque sus actividades estén perseguidas por la ley (por ejemplo, consumir drogas ilegales), bien porque sean sancionadas socialmente (por ejemplo, la prostitución) o posean características que pueden ser estigmatizadoras (por ejemplo, estar infectado por el virus de la inmunodeficiencia humana (VIH) o ser inmigrante ilegal). Por todo ello, a estas poblaciones no se les pueden aplicar los métodos de muestreo tradicionales puesto que no existe un marco muestral o su construcción es extremadamente costosa y poco práctica debido a su naturaleza. Los investigadores con recursos limitados pueden inclinarse a utilizar muestras de conveniencia ad hoc cuando se enfrentan a una población especial o rara. Si bien estas muestras de conveniencia pueden ser adecuadas para una etapa exploratoria de la investigación, son en realidad inadecuadas para hacer estimaciones sobre la población oculta.

Los métodos de muestreo en poblaciones ocultas tienen como objetivo idear medios para extraer muestras que sean representativas o que cubran la heterogeneidad de la población objetivo [3]. Existen algunos métodos no probabilísticos disponibles que tienen en cuenta grupos pequeños dentro de una población, como el muestreo en bola de nieve o el muestreo dirigido por los participantes (RDS por sus siglas en inglés: Respondent-driven sampling). RDS es un método de muestreo tipo bola de nieve que se utiliza para estudiar poblaciones ocultas. Estas son poblaciones que carecen de un marco de muestreo confiable y, por lo tanto, son de difícil acceso. RDS fue introducido por primera vez por Heckathorn [4] y luego fue desarrollado por Salganik y Heckathorn [5] y Volz y Heckathorn [6]. RDS hace uso de las redes sociales de los miembros de la población. El proceso de selección comienza con un conjunto de miembros iniciales de la población objetivo, seleccionados por conveniencia, llamados semillas. Estos encuestados reciben cupones de reclutamiento (generalmente tres), de modo que reclutan a la próxima ola de participantes entre sus contactos conocidos dentro del grupo oculto, generalmente con incentivos [4]. Cuando estos encuestados devuelven sus cupones, reclutan a la siguiente ola de participantes. Este proceso continúa hasta que se alcanza el tamaño de muestra deseado [5]. Algunos ejemplos populares de RDS son:

- Personas en riesgo de contraer el VIH [7] y usuarios de drogas inyectables [8].
- La comunidad LGBTI [9].
- Minorías étnicas e inmigrantes [10, 11].
- Niños de la calle [12].

RDS no requiere un marco de muestreo ordinario y reduce los costos involucrados en comparación con el muestreo tradicional.

1.1. Algunos métodos de muestreo en poblaciones ocultas

La metodología para la inferencia basada en datos RDS surge a partir del esfuerzo de la comunidad científica para estudiar poblaciones ocultas. A continuación, se describen algunos métodos de muestreo en poblaciones ocultas, para luego en el siguiente capítulo introducir formalmente el muestreo dirigido por los participantes y desarrollar los estimadores RDS más usuales.

1.1.1. Muestreo de referencia en cadena

Muestreo en bola de nieve

Para muestrear poblaciones ocultas, el muestreo de referencia de cadena es adecuado cuando los miembros de la población objetivo se conocen entre sí y están estrechamente interconectados. El muestreo comienza con un conjunto de sujetos iniciales que sirven como semillas para una cadena de referencias en expansión, con sujetos de cada ola que remiten a los sujetos de la ola siguiente. El muestreo en bola de nieve fue introducido por Goodman [13] - Este muestreo se inicia con una muestra de conveniencia accesible; estos individuos proporcionan los nombres de un número fijo de otras personas que cumplen los criterios de investigación. El investigador se acerca a estas personas y se pide a cada sujeto que está de acuerdo que proporcione una cantidad fija de nombres adicionales. El investigador continúa este proceso durante tantas etapas como desee.

El muestreo en bola de nieve junto con otros métodos de muestreo, presentan varios problemas descritos por Erickson [14] que indicamos a continuación:

- Las inferencias sobre los individuos deben basarse principalmente en la muestra inicial (muestra de conveniencia), ya que los individuos adicionales encontrados por las referencias en cadena nunca se encuentran al azar.
- Las muestras de referencia en cadena tienden a estar sesgadas hacia los sujetos más cooperativos que aceptan participar; este problema se agrava cuando los sujetos iniciales son voluntarios, porque en términos de cooperación pueden ser atípicos.
- La muestra puede estar sesgada debido al enmascaramiento, esto es otro tipo de problema, que ocurre cuando los encuestados no revelan la identidad o ubicación de otra persona (por ejemplo, para un estudio del uso de drogas ilícitas). Esto puede ser un problema importante cuando una población tiene fuertes preocupaciones de privacidad (poblaciones ocultas).
- Las referencias se producen a través de enlaces de red, de modo que los sujetos con redes personales más grandes serán sobremuestreados y se excluirán los aislados relativos.

Debido a estos cuatro puntos, se suelen considerar que las muestras en bolas de nieve carecen de cualquier afirmación válida para producir muestras representativas y consistentes.

Se ha desarrollado el muestreo del informante clave y el muestreo específico para superar las dificultades que aquejan a las muestras en bola de nieve.

Muestreo del informante clave

El muestreo del informante clave propuesto por Deaux y Callaghan [15] está diseñado para superar los sesgos de respuesta, seleccionando encuestados especialmente informados y preguntándoles sobre el comportamiento de los demás, en lugar del propio. Este método reduce la tendencia a exagerar el comportamiento socialmente aceptable y subestima el comportamiento de mala reputación, sin embargo, agrega varias fuentes de sesgo que listamos a continuación:

- Cuando los profesionales son informantes clave, su orientación profesional puede sesgar sus respuestas.
- Los informantes clave pueden carecer de un conocimiento suficientemente detallado; por ejemplo, el conocimiento sobre los ingresos diarios de los trabajadores informales es estrictamente personal.
- Los informantes clave no interactúan con un grupo aleatorio de la población de estudio. Si el informante clave es un profesional, el sesgo es una forma de sesgo institucional característica de las muestras extraídas de poblaciones institucionalizadas. Por ejemplo, los individuos que no se autoidentifican públicamente como parte de la población no son seleccionados al azar.

Por lo tanto, el enfoque de informante clave tiene limitaciones: introduce nuevas fuentes de posibles sesgos de respuesta, no puede utilizarse para acceder a información personal y altamente detallada y el muestreo puede tener un sesgo institucional.

1.1.2. Muestreo específico

El muestreo específico formulado por Watters y Biernacki [16] es una respuesta ampliamente empleada a las deficiencias de los modelos de referencia de cadena y es adecuado cuando la población objetivo se encuentra geográficamente concentrada [17]. Implica dos pasos básicos:

- Los investigadores de campo mapean una población objetivo. En la medida en que se logra penetrar en las redes locales que vinculan a los posibles encuestados, se puede evitar el submuestreo.

- Los investigadores de campo reclutan un número de sujetos previamente especificado en sitios identificados por el mapeo etnográfico, asegurando que los sujetos de diferentes áreas y subgrupos aparecerán en la muestra final.

La adecuación del muestreo específico depende de la precisión y la exhaustividad del mapeo etnográfico. Si el mapeo etnográfico es casi perfecto, el muestreo específico se reduce a una forma de muestreo específico por ubicación (o target/location sampling). Sin embargo, el muestreo mediante este método siempre es limitado, ya sea por los efectos de la hora del día en que los investigadores reclutan, dónde realizan su reclutamiento y las estrategias de reclutamiento que utilizan. Por lo tanto, en la práctica el tiempo y los recursos disponibles para realizar evaluaciones etnográficas exhaustivas limitan la utilidad de este enfoque [4, 18].

1.1.3. Muestreo espacial

Otro enfoque que se está utilizado cada vez más en los últimos años aprovecha el hecho que algunas poblaciones ocultas tienden a reunirse en ciertos tipos de lugares [17]. Por ejemplo, las trabajadoras sexuales a menudo se congregan en burdeles, salones de masajes y zonas de tolerancia. En el muestreo espacial (TLS por sus siglas en inglés: time-location sampling) introducido por Muhib et al. [19], dichos sitios se enumeran en un mapeo etnográfico preliminar o en un ejercicio de evaluación previa a la vigilancia; la lista de sitios así desarrollada se utiliza como un marco de muestreo del cual elegir una muestra probabilística de sitios y los datos se recopilan de todos o de una muestra de los miembros del subgrupo que se encuentran en el sitio durante un intervalo de tiempo predefinido (por ejemplo, un período de tiempo de 3 horas en un día de la semana elegidos ambos al azar). Dado que se pueden calcular las probabilidades de selección, TLS califica como un método de muestreo probabilístico.

Sin embargo, a menos que se identifiquen todos o un porcentaje muy alto de los sitios donde se congregan los miembros del subgrupo para que puedan ser incluidos en el marco de muestreo y todos o un porcentaje muy alto de los miembros del subgrupo visiten dichos sitios al menos periódicamente, TLS también sufre de niveles inaceptables de sesgo. En teoría, la inclusión de todos los sitios de recolección se puede lograr con el tiempo y los recursos suficientes para el desarrollo del marco de muestreo, pero aquí nuevamente existen límites prácticos en cuanto a los recursos que pueden destinarse a tales actividades de manera regular. Debido a que los lugares donde se congregan los miembros de subgrupos particulares

cambian con el tiempo, es necesario repetir el ejercicio de desarrollo del marco de muestreo antes de cada ronda de recopilación de datos de vigilancia. Tener disponible el marco de muestreo de rondas de vigilancia anteriores reduce el costo del desarrollo del marco de muestreo en rondas posteriores, pero los costos de actualizar el marco de muestreo tienden, no obstante, a no ser triviales. Como resultado, existe un peligro real de perder algunos sitios, lo que resulta en un posible sesgo de muestreo.

Los miembros del subgrupo que no visitan dichos sitios plantean un problema más grave. Aquí, ninguna cantidad de rigor en la construcción de marcos de muestreo de los sitios de recolección reducirá el sesgo de muestreo. Por lo tanto, si una proporción significativa de miembros de un subgrupo dado tienden a no frecuentar dichos sitios, TLS puede estar sujeto a un sesgo de muestreo grave, en la medida en que las características de interés científico de los miembros del subgrupo que no visitan los lugares de reunión difieren de los que sí lo hacen.

Otra fuente importante de sesgo con TLS es la naturaleza de los sitios de reclutamiento. Los usuarios de drogas inyectables que vengan a comprar drogas querrán irse lo antes posible. Las trabajadoras sexuales que trabajan en una esquina no querrán perder a un cliente potencial. La falta de respuesta estará estrechamente relacionada con ciertos sitios.

Con el objetivo de superar las limitaciones inherentes a los métodos descritos anteriormente se ha producido un renovado interés en la bola de nieve y otros métodos de referencia en cadena. Estos métodos tienen un gran potencial, como revelan los estudios de redes sociales [20]. El enfoque más actual es el muestreo dirigido por los participantes o RDS que se describe en la siguiente sección.

1.2. Introducción al muestreo dirigido por los participantes

El enfoque más nuevo para el muestreo de poblaciones ocultas se conoce como muestreo dirigido por los participantes o RDS. El método es similar al muestreo en bola de nieve en el sentido que implica un muestreo de referencia en cadena. Sin embargo, el proceso de reclutamiento se implementa de una manera que permite el cálculo de las probabilidades de selección, haciendo que el tamaño de la red personal de cada participante haga el papel de probabilidad de inclusión. Además, el método no se limita a los miembros del subgrupo que son accesibles en los sitios, sino que extiende la muestra a todos los miembros potenciales de un subgrupo seleccionado, accediendo a los encuestados a través de sus redes sociales [4]. Es un enfoque para el estudio de poblaciones ocultas o de difícil acceso. Las poblaciones

difíciles de alcanzar se caracterizan por la dificultad de muestrearlas utilizando métodos de probabilidad estándar. Por lo general, no está disponible un marco de muestreo para la población objetivo, la cual suele ser rara y en ocasiones estigmatizada, por lo que es prohibitivamente costoso contactar a los participantes a través de los marcos disponibles. RDS presenta dos características principales para este entorno: un diseño para el muestreo de la población objetivo y un método para realizar inferencia de estas poblaciones en base a la muestra resultante. De la primera característica es de donde el método toma su nombre y este muestreo se basa en tres puntos:

- Selección de semillas: Todos los estudios de RDS comienzan con una pequeña cantidad de semillas de la población objetivo que corresponden a la ola 0. Las semillas deben ser diversas y estar bien interconectadas, pero no es necesario elegirlas al azar.

- Entrevistas y reclutamiento: Luego que las semillas completan el proceso de entrevistas, reciben un número predeterminado de cupones (generalmente 2 o 3) que pueden usar para reclutar a otras personas que formarán parte de la ola 1. Los reclutados de la ola 1, posteriormente completan el proceso de entrevista y reclutan a otros individuos en la ola 2. Esta cadena de referencias (que es formada por los reclutados por cada semilla en sucesivas olas) se inicia en la ola 0 y continúa hasta que se alcanza el tamaño de muestra deseado.

- Incentivos: Los participantes reciben dos incentivos: uno por completar la entrevista y otro por cada compañero que se recluta con éxito (en promedio 14 dólares [18]).

Este proceso de reclutamiento puede ser modelado como un proceso de Markov, una forma de proceso estocástico con dos características esenciales. Primero, el proceso asume un número limitado de estados. En segundo lugar, el proceso depende del estado, donde la probabilidad de pasar de un estado a otro depende de una matriz de probabilidad de transición.

El proceso de reclutamiento tiene varias características. Primero, es un proceso sin memoria, en el que los patrones de reclutamiento dependen solo del reclutador, no del reclutador del reclutador. Esto significa que el reclutamiento corresponde a lo que se denomina un proceso de Markov de primer orden. Segundo, no hay grupos reclutados exclusivamente desde adentro (es decir, la homofilia[1] es siempre menor a 1). Por

[1]La homofilia es la tendencia a asociarse con personas con características similares, se calcula como la relación entre el número de reclutados con la misma característica que su reclutador y el número esperado por azar [21]. Un valor de 1 significa que no hay reclutamiento preferencial, mientras que los valores superiores a 1 indican homofilia y los valores inferiores a 1 heterofilia.

lo tanto, el reclutamiento es ergódico. Un proceso se denomina ergódico si, a medida que un proceso se mueve de un estado a otro, cualquier estado puede repetirse y la probabilidad de que un estado se repita es mayor que cero. En una aplicación concreta los estados se refieren a las características de los sujetos, el movimiento de estado a estado se refiere a un reclutador con un conjunto de características que recluta a otro sujeto con las mismas o diferentes características y que cualquier estado puede repetirse significa que después de una o más oleadas de reclutamiento, un reclutado puede tener las mismas características que el reclutador anterior. En esencia, esto significa que el reclutamiento no puede quedar atrapado dentro de un solo grupo o conjunto de grupos, como ocurre cuando la cadena de reclutamiento ingresara a ese grupo y no es posible la salida (es decir, ningún reclutamiento externo). Por lo tanto, el reclutamiento corresponde a lo que se denomina un proceso de Markov regular. Además, la estrategia concebida de esta manera reduce las preocupaciones de confidencialidad generalmente asociadas al muestrear poblaciones de individuos estigmatizados.

La segunda característica principal es la estimación de los parámetros poblacionales. Al igual que con la mayoría de los estudios de poblaciones ocultas, una muestra RDS comienza con una muestra de conveniencia de individuos. La característica clave es que, según la teoría de la cadena de Markov, este procedimiento produce muestras que son independientes de los sujetos iniciales a partir de los cuales comienza el muestreo (no importa si la muestra inicial se toma o no de forma aleatoria), reduce los sesgos resultantes del voluntarismo, el enmascaramiento y controla los sesgos que resultan de las diferencias en los tamaños de las redes personales.

RDS aborda una necesidad desatendida y esta se demuestra en la explosión de estudios de RDS en el mundo (Europa, Asia, América, África y Australia). Abdesselam et al. [22] identifican un total de 932 artículos, donde 900 aplican la metodología RDS en poblaciones de usuarios de drogas, hombres que tienen sexo con hombres, trabajadoras sexuales, migrantes y minorías étnicas. Mientras que los 32 restantes describen estrategias para mejorar la inferencia de RDS.

1.3. Objetivos

La mayoría de los trabajos que aplican la metodología RDS se centran en poblaciones de usuarios de drogas, hombres que tienen sexo con hombres, trabajadoras sexuales y migrantes, pero apenas hay trabajos realizados para muestrear poblaciones de minorías étnicas. Entonces, estudiar por ejemplo a una

población indígena de Sudamérica puede llenar un vacío ya que no se ha realizado un enfoque de este tipo para estudiar a este grupo étnico.

Por otro lado, la literatura actual de datos RDS carece de enfoques basados en principios para el modelado multivariable [23]. En particular los enfoques existentes en regresión en el contexto de RDS son tal que: i) únicamente se centra en la estimación sin tomar en cuenta la estructura de la red RDS [24]; ii) incluyen ponderaciones de RDS en el modelo de regresión, ignorando la dependencia entre los datos dentro de la red RDS [25]; iii) incorporan las semillas como efectos aleatorios para ajustar la dependencia entre datos dentro de la red RDS, omitiendo las ponderaciones de RDS [26]; iv) integran efectos mixtos que incluye efectos aleatorios sobre características como semillas y reclutadores para dar cuenta de la dependencia y el uso de ponderaciones en diferentes niveles de agrupación. Además, modela los efectos sociales impulsados por la homofilia al incluir un parámetro para tener en cuenta las posibles interacciones entre los reclutadores y los valores de los reclutados de las covariables homófilas [27] y v) incorporan efectos mixtos generalizados, con efectos impulsados por la homofilia para tratar las covariables homofílicas y con efectos aleatorios espaciales para modelar la dependencia entre los resultados dentro de la red y la inclusión de ponderaciones RDS para tener en cuenta el muestreo no aleatorio de la población objetivo, cuando los individuos reclutados informan con precisión el tamaño de su red personal [28]. Por lo tanto, existen trabajos muy interesantes basados en la modelización que intentan considerar la agrupación de redes subyacentes a la muestra y la homofilia inherente al proceso de reclutamiento. Sin embargo, no se tiene un método estandarizado para el modelado de regresión utilizando datos recopilados a través de RDS [29, 30]. Entonces, debido a que comprender la dependencia entre variables es a menudo un objetivo principal en la investigación, es importante abordar el problema del modelado de regresión y la asociación entre variables con datos RDS.

Los objetivos en los cuales se centra el trabajo son los siguientes:

1. Poner en práctica una encuesta no probabilística mediante la metodología RDS, para recoger información sobre poblaciones indígenas y otras minorías étnicas.

2. Formular expresiones de estimadores de covarianza y coeficientes de correlación en el contexto del muestreo de poblaciones difíciles de alcanzar.

En este contexto presentamos la revisión de los métodos del muestreo dirigido por los participantes o RDS, organizado de forma sistemática, unificando la notación y resaltando sobre el estimador RDS

la información empleada, sus limitaciones, aportaciones distintivas y tipo de estimación de la varianza. También abordamos el problema de encuestar a una población de indígenas y otras minorías étnicas realizando una encuesta RDS en el cantón Riobamba-Ecuador y proponemos un nuevo método de estimación ponderada para datos continuos en el problema del modelado de regresión y la asociación entre variables.

Se comienza en el siguiente capítulo en su primera sección describiendo los métodos del muestreo dirigido por los participantes, empezando con una notación y conceptos básicos, seguido se presenta de forma detallada las dos características principales de la metodología RDS. Los modelos y estimadores RDS más utilizados en la literatura se consideran en la siguiente sección y dentro de esta empezamos describiendo el modelo RDS como un proceso de Markov, luego se presenta el estimador de Volz y Heckathorn [6] basado en la red social. Se sigue con el estimador de Salganik y Heckathorn [5] basado en la reciprocidad. Se exponen los métodos de estimación de la varianza como son el método Bootstrap y el basado en la cadena de Markov Monte Carlo (MCMC). El estimador de Gile y Handcock [31] basado en el muestreo sucesivo se presenta en la sección siguiente. Finalmente se presentan otros enfoques de la metodología RDS.

En el tercer capítulo se describe la metodología RDS de encuestas en minorías étnicas. En la primera sección se expone la necesidad de información sobre minorías étnicas, seguido se muestran los retos que representa la recogida de información y el análisis de datos en encuestas RDS. En la última sección se presenta una contribución original que trata sobre el problema de recolectar información sobre minorías étnicas en Ecuador mediante la metodología RDS.

El capítulo cuatro está destinado a la presentación de la regresión con datos RDS. En la primera sección se exponen las preocupaciones en la regresión con datos RDS y se describe el estudio previo al modelado de regresión. La siguiente sección describe dos estrategias para realizar un análisis de regresión con datos RDS. La última sección presenta una de las aportaciones fundamentales del presente trabajo, que es el estimador ponderado de la muestra para datos continuos en el problema del modelado de regresión y la asociación entre variables continuas.

Capítulo 2

Muestreo dirigido por los participantes

La metodología para la inferencia basada en datos RDS se ha desarrollado lentamente. El artículo original de Heckathorn [4] establece supuestos sólidos sobre el procedimiento de muestreo para asegurarse que las muestras seleccionadas por RDS son muestras representativas de la población objetivo de estudio. Salganik y Heckathorn [5] introducen el argumento de la Cadena de Markov para la mezcla de la población y proponen un estimador basado en igualar el número de vínculos cruzados entre dos subpoblaciones de interés. Luego Volz y Heckathorn [6] introducen el estimador análogo al de Hansen-Hurvitz, que se basa en gran medida en la red social. Más actual es el estimador de Gile [31] basado en el muestreo sucesivo.

Se comienza en la sección 2.1 presentando los desafíos de la recopilación y el análisis de datos en RDS. En la sección 2.2, se presenta la notación y conceptos básicos del muestreo dirigido por los participantes y en la sección 2.3, se describen los modelos y estimadores más utilizados en la literatura de RDS.

2.1. Recopilación y análisis de datos

El muestreo dirigido por los participantes, ha tenido un gran éxito en la adquisición de datos confiables sobre poblaciones de difícil acceso, principalmente para la investigación de salud pública. Además, RDS se ha utilizado eficazmente para muestrear a poblaciones de minorías étnicas [57, 58, 59, 60]. En esta

sección, siguiendo a Tyldum y Johnston [61] se describen los desafíos de la recopilación y el análisis de datos en RDS.

2.1.1. Estructura de la población objetivo

El muestreo dirigido por los participantes se emplea en encuestas para recopilar todo tipo de información, aunque solo es necesario recopilar la siguiente información: las conexiones entre los reclutados y los reclutadores, que se rastrea mediante el uso de cupones y el tamaño de la red personal (TRP) del encuestado, que mide cuántos individuos de la población objetivo conocen los encuestados y quiénes también los conocen [4]. El TRP de cada encuestado se mide mediante una pregunta abierta o una serie de preguntas. Construir la(s) pregunta(s) y recopilar respuestas precisas de los encuestados no es sencillo. Dados los desafíos para obtener respuestas a la(s) pregunta(s), se utilizan técnicas de estimulación y a menudo, se coloca(n) la(s) pregunta(s) cerca del comienzo de la encuesta. Se necesitan cuatro elementos al construir la(s) pregunta(s) del TRP: i) una definición clara de la población objetivo (y, por lo tanto, los criterios de elegibilidad); ii) un entendimiento compartido o una definición explícita de lo que es "conocer" a alguien; iii) un límite geográfico para la encuesta y iv) un período de referencia claramente definido durante el cual el encuestado ha entrado en contacto con los pares informados.

2.1.2. Selección de semillas y su función

Se debe prestar especial atención a la selección estratégica de semillas, su formación y función en el proceso de muestreo RDS. Aunque la selección de semillas se lleva a cabo al comienzo de una encuesta, este proceso necesita una consideración cuidadosa, ya que es probable que afecte la recopilación y el análisis de datos.

Para seleccionar estratégicamente las semillas, los investigadores deben conocer su población objetivo a fin de superar los posibles cuellos de botella[1] y lograr una muestra compuesta por una mezcla diversa de reclutados. En el mismo sentido, la selección estratégica del número "correcto"de semillas resulta crucial para completar con éxito una encuesta RDS. Idealmente, estas personas deben tener grandes redes sociales compuestas por personas diversas, pueden superar los cuellos de botella y reclutar a otras personas que participarán en la encuesta.

[1]En la terminología de RDS, las partes de una población que están más densamente conectadas que otras partes se denominan conglomerados y cuando hay pocas conexiones entre subgrupos particulares se denominan cuellos de botella.

2.1.3. Determinación y distribución de incentivos

El doble incentivo es un componente central de la metodología RDS, donde los encuestados reciben un incentivo por participar en la encuesta e incentivos adicionales para contratar nuevos encuestados. Los incentivos económicos se utilizan cada vez más en las encuestas ordinarias para impulsar la participación, pero en RDS dichos incentivos[2] tienen funciones adicionales y, en muchos sentidos, son parte integral de la metodología de la encuesta.

Las diferentes poblaciones plantean distintos desafíos para determinar el mejor incentivo para usar. Por ejemplo, en estudio de inmigrantes aquellos grupos más ricos o con más educación pueden no estar motivados por los modestos incentivos económicos que se utilizan comúnmente en RDS, pero pueden responder bien a incentivos alternativos, presión de grupo, deseo de ser escuchado o de contribuir a la investigación. Los encuestados con ingresos más bajos pueden responder bien a incentivos financieros modestos, pero no hay garantía que el incentivo planificado funcione bien. Las señales que los incentivos son demasiado altos incluyen intentos de inscripción por parte de personas que no son miembros de la población objetivo y la venta y el trueque de cupones. Los incentivos que son demasiado bajos pueden resultar en un reclutamiento lento o nulo. Se pueden reducir los incentivos elevados si se puede evitar el resentimiento en la población de interés. Los incentivos bajos pueden aumentarse, si el presupuesto lo permite, o pueden complementarse con incentivos alternativos no materiales. Sin embargo, es importante obtener el incentivo desde el principio o cambiarlo lo antes posible si es necesario, ya que el cambio de incentivos también afecta las probabilidades de reclutamiento.

2.1.4. Modificaciones en la recopilación de datos

Como en toda investigación de encuestas, la recopilación de datos RDS es la culminación de meses de planificación cuidadosa. No obstante, puede ser difícil predecir cómo se desarrollará una encuesta RDS, especialmente cuando el método no se ha utilizado antes en una población en particular. Por lo general, no sabemos de antemano hasta qué punto nuestra población objetivo responderá al método de muestreo, o hasta qué punto las características de la encuesta, como los incentivos y las horas de operación, se adaptarán a las necesidades de nuestros encuestados. Por lo tanto, hacer que RDS funcione a menudo implica un proceso iterativo de observar qué tan bien funciona el método y estar preparado para modificar

[2]Los incentivos monetarios son los más utilizados, pero algunas encuestas se basan en incentivos alternativos, como tarjetas de regalo u otros pequeños artículos de valor relevantes para la población de estudio.

ciertos elementos para adaptarlo mejor a la población.

Muchas encuestas RDS han tenido que modificar algunas características de diseño durante la recopilación de datos. Siempre que se cumplan los rigurosos principios metodológicos de RDS, son posibles muchas modificaciones durante la etapa de recopilación de datos. Sin embargo, ciertos aspectos de la recopilación de datos de RDS no deben modificarse una vez que haya comenzado la encuesta. Esto incluye los criterios de elegibilidad sobre los cuales se determina la probabilidad de selección a través de los tamaños de la red personal. La logística de la encuesta, por otro lado, puede y a menudo debe adaptarse para garantizar el éxito de la contratación en RDS. La evaluación formativa y el monitoreo paralelo continuo son particularmente importantes para identificar desafíos y adaptarse a la población lo más rápido y de la mejor manera posible.

2.1.5. Análisis de datos

El muestreo dirigido por los participantes es una metodología para recopilar y analizar datos. Si los datos recopilados mediante RDS no se analizan para corregir sesgos específicos, la encuesta no debe denominarse RDS (a menos que se pueda demostrar que los datos son completamente autoponderados, es decir, las estimaciones ajustadas son las mismas que las estadísticas no ajustadas). Para analizar los datos RDS, es necesario utilizar uno de los programas de software especializados que generan estimadores e intervalos de confianza específicos para los supuestos de RDS. Se indican a continuación los software disponibles de nuestro conocimiento con sus sitios web:

- RDSAT (www.respondentdrivensampling.org): Este sitio web es mantenido por Douglas Heckathorn, el padre de RDS. Alberga un enlace para descargar RDSAT, así como el Manual de usuario de RDSAT. El sitio web también contiene algunos antecedentes generales sobre RDS y enlaces a "referencias básicas de RDS".

- RDS Analyst (http://hpmrg.org/): Este sitio web es mantenido por el Grupo de Investigación de Métodos de Población de Difícil Acceso, que incluye a varios investigadores destacados de RDS. Alberga un enlace para descargar el programa RDS Analyst, así como numerosos tutoriales y guías sobre la instalación y el uso del software. RDS Analyst es un sistema de análisis de datos gráfico intuitivo y multiplataforma para el análisis de datos RDS. Utiliza menús y cuadros de diálogo para guiar al usuario de manera eficiente a través del proceso de manipulación y análisis de datos y tiene

una hoja de cálculo similar a Excel para facilitar la visualización y edición del marco de datos.

- Paquete RDS (https://cran.r-project.org/web/packages/RDS/index.html): Es una implementación en software R que proporciona funcionalidad para realizar estimaciones con datos RDS. Esto incluye los estimadores RDS I y RDS II de Heckathorn, así como el estimador de muestreo sucesivo de Gile. El paquete RDS es parte de RDS Analyst.

Evaluación de sesgos

En el análisis RDS hay varias formas de evaluar el nivel de sesgo. El sesgo en los datos RDS es específico de las variables analizadas; puede haber un mayor sesgo en una variable en comparación con otra variable de la muestra. Por ejemplo, si bien el sesgo en las variables de sexo o edad puede ser grande, el sesgo en la variable de educación puede ser pequeño. A continuación discutimos cuatro posibles fuentes de sesgo en RDS: dependencia de semillas, el grado de homofilia, actividad de reclutamiento diferencial y cuellos de botella.

Dependencia de semillas: El muestreo RDS comienza con semillas seleccionadas intencionalmente, que pueden tener o no características que representen la estructura de red subyacente de la población. Los individuos tienden a ser similares a los demás en sus redes sociales en una serie de características, como nivel educativo, lugar de residencia y preferencias políticas. Por lo tanto, si la muestra no llega a todas las subpoblaciones de una red, lo cual es probable si las cadenas de reclutamiento son cortas, entonces la muestra puede representar las características de las semillas en lugar de las características de la población. Una forma de determinar si la muestra final depende de las semillas es medir si se ha alcanzado el equilibrio o la convergencia en la muestra. El equilibrio se evalúa determinando la proporción de la muestra (es decir, dividiendo el número de personas con o sin una característica sobre el número total de personas de la muestra) en cada oleada sucesiva. En algún momento (en una ola en particular), la proporción de muestra ya no cambiará de una ola a otra. Este punto de equilibrio indica que la muestra ha comenzado a representar una combinación aleatoria de características sobre las que se estructura la población. El logro del equilibrio no es el punto en el que se debe detener el muestreo. Más bien, es necesario reclutar muchas olas más allá del punto de equilibrio para asegurar que el equilibrio de proporciones permanezca estable durante numerosas olas sucesivas y para lograr un tamaño de muestra adecuado que no esté marcado por la dependencia de las semillas. El punto en el que se alcanza el equilibrio puede ser diferente para cada

variable de la misma muestra. Es importante señalar que lograr el equilibrio no es una indicación que la muestra esté completamente libre de sesgos.

Homofilia: La homofilia de reclutamiento se calcula como la razón del número de reclutados que comparten la misma característica que su reclutador; relativo al número que se espera si el reclutamiento fuera aleatorio. Una alta homofilia de reclutamiento puede ser un indicativo que existe un sesgo de semilla (es decir, que no se ha alcanzado el equilibrio) y/o que el reclutamiento está estancado en un subgrupo (por ejemplo, solo los hombres están representados en una muestra de hombres y mujeres). Un homofilia alta dará como resultado estimaciones inestables y una mayor varianza.

Actividad de reclutamiento diferencial: La actividad de reclutamiento diferencial es un indicativo de la conexión relativa de un grupo a otro. Se mide como la relación entre el tamaño medio de las redes sociales personales de un grupo (por ejemplo, mujeres) en relación con el tamaño medio de las redes sociales personales de otro grupo (por ejemplo, hombres). El reclutamiento diferencial da como resultado una representación insuficiente o excesiva de algunos grupos en una muestra RDS.

Cuellos de botella: Un cuello de botella es un indicativo de la existencia de pocas conexiones entre subgrupos particulares de una población. Los cuellos de botella son una forma extrema de homofilia [62] que puede aumentar sustancialmente el efecto de la selección inicial de las semillas en las estimaciones [63].

2.2. Notación y conceptos básicos

Los datos de la red social generalmente consisten en un conjunto de N vértices (comúnmente conocidos como nodos) y una variable de enlace relacional, Z_{ij}, medida en cada par de vértices ordenados posibles, $\{i,j\}, i,j = 1,\ldots,N, i \neq j$ que llamaremos arcos. En los casos más simples Z_{ij} es una variable dicotómica que indica la presencia o ausencia de alguna relación de interés, como la amistad, la colaboración, la transmisión de información o la enfermedad, etc. Esto se representa mediante una sociomatriz Z de tamaño $N \times N$ con elementos diagonales tratados como ceros. En el caso de las relaciones binarias, los datos también se pueden considerar como un grafo en el que los vértices son actores y el conjunto de arcos está expresado por $\{(i,j) : Z_{ij} = 1\}$. Para muchas redes, las relaciones no están dirigidas en el sentido que se cumple $\{Z_{ij} = Z_{ji}, i,j = 1,\ldots,N,\}$ y nos referimos a estos como aristas. A menudo variables asociadas a las aristas, son las características de interés científico de la población que se representa por el vector y

de tamaño N.

Gile [32] muestra que dentro del análisis de redes se consideran dos perspectivas, una referente al análisis descriptivo o de población finita y otra sobre la inferencia del mecanismos generador o un marco de superpoblación. En este sentido, en el marco de la superpoblación como realización de variables aleatorias, la estructura relacional Z es una matriz aleatoria, mientras que, en el marco de las poblaciones finitas, la estructura relacional es fija[3]. Enfatizamos esta distinción utilizando z para representar la estructura relacional en el marco de una población finita. Así el grado del vértice del individuo i lo notamos mediante $\delta_i = \sum_j z_{ij}$.

En la inferencia basada en el diseño para datos de red, la población de valores se trata como fija y toda la incertidumbre en las estimaciones de las características de la población no observada se debe al mecanismo de muestreo, que habitualmente se asume completamente conocido.

La inferencia basada en el diseño generalmente se enfoca en obtener estimadores insesgados para estimar parámetros de interés de la población. La mayor parte del trabajo sobre la estimación basada en el diseño para redes muestreadas se enfoca en estimar dos cantidades:

1. La suma $\tau = \sum_{i\in\{1,\dots,N\}} y_i$ de una variable nodal y en todos los nodos de la población.
2. El número total de vínculos en la red. En una red no dirigida, este es el número de aristas $N_E = \sum_{i<j,i,j\in\{1,\dots,N\}} z_{ij}$ y en una red dirigida es el número total de arcos $N_A = \sum_{i,j\in\{1,\dots,N\}} z_{ij}$.

Tenga en cuenta que τ es una propiedad de las variables nodales, mientras que N_E y N_A son propiedades de la estructura relacional.

En el mismo sentido, se puede estar interesado en estimar la media poblacional $\mu = \tau/N$, esto se logra mediante el estimador de Horvitz Thompson [33] que es una herramienta clásica de inferencia basada en el diseño. Consiste en una suma ponderada de probabilidad inversa de las observaciones en la muestra. Si conociéramos la probabilidad de muestreo nodal π_i de cada individuo i muestreado, podríamos estimar $\mu = \frac{1}{N}\sum_{i\in\{1,\dots,N\}} y_i$ de la siguiente forma para el caso de muestreo sin reemplazo:

$$\hat{\mu} = \frac{1}{N}\sum_{i:s_i=1}\frac{y_i}{\pi_i}, \tag{2.1}$$

donde

[3]En el marco de las poblaciones finitas se pueden introducir modelos de superpoblación, contemplando que los residuos pueden tener dependencia.

- s es un vector aleatorio binario de tamaño n (número de elementos en la muestra) que indica el subconjunto de vértices muestreados, en el cual el elemento i-ésimo es 1 si el vértice i-ésimo es parte de la muestra y 0 en caso contrario.

- Si el individuo i se muestrea con probabilidad π_i entonces, $\pi_i > 0 \; \forall \; i \in U$ con $U = \{1, \ldots, N\}$.

Hay dos dificultades con este enfoque en el contexto de RDS: el tamaño de la población N es desconocido y las probabilidades de inclusión π_i son desconocidas. El primero es fácilmente prescindible. Utilizando un estimador insesgado de N, $\hat{N} = \sum_{i:s_i=1} \frac{1}{\pi_i}$, se obtiene:

$$\hat{\mu}^* = \frac{\sum\limits_{i:s_i=1} \frac{y_i}{\pi_i}}{\sum\limits_{i:s_i=1} \frac{1}{\pi_i}}, \tag{2.2}$$

que es un estimador consistente de μ. Esta es una variante estándar en el estimador clásico de Horvitz-Thompson, conocido como el estimador de Hájek. Thompson [34] sugiere que este estimador generalmente supera al estimador estándar de Horvitz-Thompson cuando las probabilidades de inclusión están lejos de ser proporcionales a y_i. Más difícil es la estimación de la probabilidad de muestreo π_i.

El diseño de muestreo de RDS se centra en la selección de una muestra considerable de individuos, al tiempo que permite la estimación de las probabilidades de inclusión π_i. RDS emplea una variante de las estrategias de muestreo de rastreo de enlaces. En este caso, los individuos están representados por vértices en una red y la relación social que facilita la transferencia de cupones RDS entre pares de individuos está representada por vínculos (aristas). En el contexto de las poblaciones difíciles de alcanzar, tales estrategias a menudo se denominan muestras en bolas de nieve [13]. Debido a que la estructura de la red social a menudo vincula a los miembros de la población objetivo, los individuos en olas anteriores pueden proporcionar acceso a otros miembros de la población. Tales muestras son, a menudo muy efectivas para alcanzar un número sustancial de miembros de la población de difícil acceso. Sin embargo y a pesar de la formulación probabilística de Goodman [13] en poblaciones difíciles de alcanzar, la ola inicial es típicamente una muestra de conveniencia, de modo que la muestra final en bola de nieve no es una muestra de probabilidad (es decir, las probabilidades de inclusión no son computables). Por lo tanto, en el muestreo en bola de nieve para poblaciones difíciles de alcanzar, la inferencia estadística válida es difícil sin supuestos adicionales.

2.3. Modelos y Estimadores

2.3.1. Modelo RDS como un proceso de Markov

Del proceso de recolección de datos RDS se obtienen las propiedades de los vértices (encuestados), así como información sobre quién recluta a quién (matriz de reclutamiento) y el tamaño de la red personal de los encuestados (grados). Esta información forma la base para generar inferencias sobre las características de la población.

Debido a la forma no aleatoria en que se recolectan las muestras de RDS, una muestra de RDS no es suficientemente representativa para la población ya que es sesgada y con diversas fuentes de sesgos, como la estructura de la red social subyacente en la que se realiza el reclutamiento y la heterogeneidad de la cadena generada por cada persona o semilla. Por ejemplo, si los individuos de la población con una determinada propiedad (por ejemplo, hombres) tienen más conexiones personales (es decir, mayor grado) que los que no tienen esta propiedad (por ejemplo, mujeres), es más probable que sean reclutados por los encuestados, lo que resulta en probabilidades de inclusión desiguales en la muestra. En consecuencia, este procedimiento de reclutamiento tiende a sobremuestrear ciertos individuos frente a otros y difícilmente puede obtener muestras representativas de la población objetivo.

Sin embargo, es posible construir modelos matemáticos para incluir pesos que permitan compensar esta sobreponderación en el reclutamiento. Estos modelos se basan en los siguientes supuestos [4, 5, 6, 18, 31, 35]:

1. Conectividad: la red en la que se realiza la contratación está conectada, es decir, todos los individuos de la población objetivo están conectados, por lo que se pueden acceder a todos a través de sus contactos personales.

2. Reciprocidad: todos los enlaces de la red no están dirigidos, es decir, las relaciones de amistad (o de conocidos) entre individuos son recíprocas: si i puede reclutar a j, entonces j puede reclutar a i.

3. El muestreo es con reemplazo: cada individuo puede participar en el estudio siempre que reciba un cupón válido, sin importar si ha participado antes.

4. Grado: los encuestados pueden informar con precisión el tamaño de sus redes personales.

5. Reclutamiento aleatorio: el reclutamiento de pares es una selección aleatoria de la red personal del

encuestado, es decir, todos los conocidos o contactos en la red personal de un reclutador tienen la misma probabilidad de recibir un cupón.

Dadas las suposiciones anteriores, si se selecciona el individuo i en la ola de muestra v, la probabilidad que cada nodo sea seleccionado en la ola $v+1$ es

$$Pr'_{i\to j} = \begin{cases} 1/\delta_i \text{ si hay una arista entre } i \text{ y } j \\ 0 \text{ en otro caso,} \end{cases} \tag{2.3}$$

y RDS se puede modelar como un proceso de Markov con la siguiente matriz de probabilidad de transición:

$$\mathbb{P} = \begin{bmatrix} 0 & z_{12}/\delta_1 & \dots & z_{1N}/\delta_1 \\ z_{21}/\delta_2 & 0 & \dots & z_{2N}/\delta_2 \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1}/\delta_N & z_{N2}/\delta_N & \dots & 0 \end{bmatrix}, \tag{2.4}$$

donde $z_{ij} = 1$ si hay un arista del individuo i al individuo j y $z_{ij} = 0$ en otro caso y δ_i es el grado de i. La distribución del estado de equilibrio para este proceso es un vector $\gamma = \{\gamma_1, \gamma_2, \dots, \gamma_N\}^T$ tal que

$$\gamma^T \mathbb{P} = \gamma^T. \tag{2.5}$$

Dado que la red no está dirigida, tenemos que $z_{ij} = z_{ji}$. Se puede verificar que (2.5) tiene como solución única

$$\gamma = \left\{ \frac{\delta_1}{\sum_{j=1}^{N} \delta_j}, \frac{\delta_2}{\sum_{j=1}^{N} \delta_j}, \dots, \frac{\delta_N}{\sum_{j=1}^{N} \delta_j} \right\}^T, \tag{2.6}$$

tal que $\sum_{i=1}^{N} \gamma_i = 1$.

En (2.6) se indica que cuando una muestra RDS alcanza el equilibrio, la probabilidad que cada nodo sea incluido en la muestra es proporcional a su grado:

$$Pr_i = \frac{\delta_i}{\sum_{j=1}^{N} \delta_j}. \tag{2.7}$$

2.3.2. Estimador RDS II

La conclusión de (2.7) es clave, ya que implica que, incluso si la muestra es seleccionada de manera no aleatoria, se puede tratar la muestra RDS como una muestra probabilística de manera que la probabilidad de inclusión de cada sujeto en la muestra se pueda aproximar por su grado y esta se puede utilizar como ponderación muestral para generar estimaciones de la población.

De forma específica, sea una muestra $s = \{s_1, s_2, \ldots, s_n\}$, donde n_A es el número de encuestados en la muestra con propiedad A (por ejemplo, un individuo autoidentificado como indígena) y $n_B = n - n_A$ es el resto de la muestra. Sea además $\{\delta_1, \delta_2, \ldots, \delta_n\}$ los grados de los encuestados.

Entonces, Pr_i puede usarse para obtener un estimador análogo al de Hansen-Hurwitz [36], en el que las observaciones se ponderan por la inversa de la probabilidad de muestreo; la proporción de individuos que pertenecen al grupo A (consideramos una propiedad binaria tal que cada individuo pertenece al grupo A o al grupo B) en la población se puede estimar mediante [6]:

$$\hat{P}_A = \frac{\sum_{i \in A \cap s} \delta_i^{-1}}{\sum_{i \in s} \delta_i^{-1}}. \tag{2.8}$$

Ahora sea la variable y alguna característica de interés real de tipo continua, como la edad o el ingreso, el estimador de la media $\hat{\bar{y}}$ toma la forma del estimador de Hájek de la siguiente manera:

$$\hat{\bar{y}} = \frac{\sum_{i \in s} \delta_i^{-1} y_i}{\sum_{i \in s} \delta_i^{-1}}, \tag{2.9}$$

con δ_i el grado informado por el entrevistado i.

El estimador que se presenta en (2.8) y (2.9) se denomina estimador RDS II (o estimador VH). El estimador RDS I (o estimador SH) apareció anteriormente en la literatura y se desarrolla en la siguiente sección. El estimador RDS II se basa en la teoría de muestreo de la cadena de Markov [37, 38] y la teoría del muestreo con probabilidades desiguales [36, 39]. Se ha demostrado que los estimadores basados en

muestras de la cadena de Markov son asintóticamente insesgados [37], es decir, que cualquier sesgo será del orden n^{-1}, donde n es el tamaño de la muestra. Por lo tanto, el estimador RDS II es asintóticamente insesgado y para tamaños de muestra significativos, cualquier sesgo es insignificante [6].

2.3.3. Estimador RDS I

El modelo recíproco

El estimador RDS I tiene una forma más complicada que RDS II y fue desarrollado en base al modelo recíproco [5]. Cuando la red no está dirigida, el número de aristas de grupos cruzados de A a B debe ser igual al número de aristas de B a A. Sea

$$C^* = \begin{bmatrix} c^*_{AA} & c^*_{AB} \\ c^*_{BA} & c^*_{BB} \end{bmatrix}, \tag{2.10}$$

la matriz de probabilidad de transición de grupos en la población, donde c^*_{XY} es la proporción de aristas del grupo X al grupo Y $(X, Y \in \{A, B\})$, tal que $c^*_{XX} + c^*_{XY} = 1$, entonces

$$N_A \bar{D}^*_A c^*_{AB} = N_B \bar{D}^*_B c^*_{BA}, \tag{2.11}$$

donde $N_A = N - N_B$ es el número de individuos del grupo A en la población y $\bar{D}^*_A$, $\bar{D}^*_B$ son los grados medios para los dos grupos.

La ecuación (2.11) se puede reescribir como:

$$P^*_A \bar{D}^*_A c^*_{AB} = (1 - P^*_A) \bar{D}^*_B c^*_{BA}, \tag{2.12}$$

donde P^*_A es la proporción de individuos del grupo A en la población.

P^*_A, $\bar{D}^*_A$, $\bar{D}^*_B$ y c^*_{XY} se estiman a partir de los datos de la muestra.

Estimación del grado medio

Dada la distribución de grados del grupo A en la red, $p_A(\delta)$, la distribución de grados de la muestra, $q_A(\delta)$, es [5]

$$q_A(\delta) = \frac{\delta p_A(\delta)}{\sum_{\delta=1}^{max(\delta)} \delta p_A(\delta)}, \tag{2.13}$$

donde $p_A(\delta)$ es la distribución de grados de la población y $\sum_{\delta=1}^{max(\delta)} \delta p_A(\delta)$ es un término de normalización para asegurar que la suma de grados $q_A(\delta)$ sea igual a 1.

Entonces, si una muestra tiene una distribución de grados, $q_A(\delta)$, entonces la distribución de grados de la población, $p_A(\delta)$, se puede estimar como

$$\hat{p}_A(\delta) = \frac{\frac{1}{\delta} q_A(\delta)}{\sum_{\delta=1}^{max(\delta)} \frac{1}{\delta} q_A(\delta)}. \tag{2.14}$$

Entonces, el grado promedio de los miembros del grupo A se puede estimar como

$$\hat{\bar{D}}_A = \sum_{\delta=1}^{max(\delta)} \delta \hat{p}_A(\delta). \tag{2.15}$$

Esto también se puede escribir como

$$\hat{\bar{D}}_A = \frac{n_A}{\sum_{i=1}^{n_A} \delta_i^{-1}}. \tag{2.16}$$

Otra forma de estimar el grado promedio es usar una razón de dos estimadores de Hansen-Hurwitz [5]: el número estimado de aristas del grupo A y el número estimado de individuos en el grupo A:

$$\hat{\bar{D}}_A = \frac{\frac{1}{n_A} \sum_{i=1}^{n_A} \frac{1}{Pr_i} \delta_i}{\frac{1}{n_A} \sum_{i=1}^{n_A} \frac{1}{Pr_i}}. \tag{2.17}$$

Reemplazando Pr_i con (2.7), tenemos

$$\hat{\bar{D}}_A = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{N}\delta_j}{\sum_{i=1}^{n_A}\frac{\sum_{j=1}^{N}\delta_j}{\delta_i}} = \frac{n_A}{\sum_{i=1}^{n_A}\delta_i^{-1}}. \tag{2.18}$$

De manera similar, el grado promedio del grupo B se puede estimar mediante

$$\hat{\bar{D}}_B = \frac{n_B}{\sum_{i=1}^{n_B}\delta_i^{-1}}. \tag{2.19}$$

Es importante observar que el numerador y el denominador de $\hat{\bar{D}}_A$ en la ecuación (2.17) son estimadores de Hansen-Hurwitz que se sabe que son insesgados [40]. También ocurre que la razón de dos estimadores insesgados resulta un estimador asintóticamente insesgado con sesgo de orden n^{-1}, donde n es el tamaño de la muestra [39]. Entonces, a medida que n_A (el tamaño de la muestra en el grupo A) aumenta, $E[\hat{\bar{D}}_A] \rightarrow \bar{D}_A^*$. Por lo general, este sesgo se considera insignificante en muestras de tamaño moderado [39].

Estimación de la matriz de probabilidad de transición de grupos

Cuando los nodos de la red se seleccionan proporcionalmente a sus grados, la probabilidad de selección de cada arista $e_{i\rightarrow j}$, se puede escribir como

$$Pr_{i\rightarrow j} = Pr_i\frac{1}{\delta_i}, \tag{2.20}$$

el término $Pr_{i\rightarrow j}$ indica la probabilidad de seleccionar al individuo j y $Pr_i\frac{1}{\delta_i}$ indica que cada arista de i tiene la misma probabilidad de ser elegido para aprobar un cupón, es decir, el supuesto 5 (reclutamiento aleatorio). Por tanto, reemplazando Pr_i con (2.7), tenemos

$$Pr_{i\rightarrow j} = \frac{\delta_i}{\sum_{j=1}^{N}\delta_j}\frac{1}{\delta_i} = \frac{1}{\sum_{j=1}^{N}\delta_j}. \tag{2.21}$$

Nótese que $\sum_{j=1}^{N}\delta_j$ es una constante para cualquier red y (2.21) indica que cuando la muestra RDS

alcanza el equilibrio, cada arista de la red tiene la misma probabilidad de ser seleccionada. En consecuencia, podemos suponer que las aristas de reclutamiento observadas en la muestra de RDS forman una muestra aleatoria de todas las aristas de la red social subyacente [5]. Existe evidencia empírica consistente con esta suposición de reclutamiento aleatorio [18]. Sea

$$C = \begin{bmatrix} c_{AA} & c_{AB} \\ c_{BA} & c_{BB} \end{bmatrix}, \tag{2.22}$$

la matriz de probabilidad de transición de grupos observada de la muestra, donde c_{XY} es la proporción de aristas del grupo X al grupo Y ($X, Y \in \{A, B\}$), tal que $c_{XX} + c_{XY} = 1$. Entonces C es una estimación asintóticamente insesgada de C^*.

Estimador RDS I

Si $\hat{\bar{D}}_A = n_A / \sum_{i=1}^{n_A} \delta_i^{-1}$ y $\hat{\bar{D}}_B = n_B / \sum_{i=1}^{n_B} \delta_i^{-1}$ son estimadores para los grados promedio del grupo A y B y si C es el estimador de la matriz de probabilidad de transición de grupos poblacional C^*, podemos entonces resolver (2.12) y obtener:

$$\hat{P}_A = \frac{c_{BA}\hat{\bar{D}}_B}{c_{AB}\hat{\bar{D}}_A + c_{BA}\hat{\bar{D}}_B}. \tag{2.23}$$

Al igual que en la ecuación (2.17), tenemos una razón de estimaciones asintóticamente insesgadas que también es asintóticamente insesgada. La ecuación (2.23) es el estimador RDS I (o estimador SH), con el que podemos hacer una estimación asintóticamente insesgada de la proporción de la población con un rasgo específico basada únicamente en los datos recopilados durante el muestreo dirigido por los participantes.

Suavizado de datos

Cuando hay más de dos grupos disjuntos en la población, el modelo recíproco genera un conjunto de ecuaciones sobredeterminadas, es decir, el número de parámetros desconocidos es menor que el número de ecuaciones. Ilustramos el suavizado de datos para tres grupos en la población, entonces el modelo recíproco se convierte en:

$$\begin{cases} 1 = \hat{P}_1 + \hat{P}_2 + \hat{P}_3 \\ \hat{P}_1 \hat{\tilde{D}}_1^* c_{12} = \hat{P}_2 \hat{\tilde{D}}_2^* c_{21} \\ \hat{P}_1 \hat{\tilde{D}}_1^* c_{13} = \hat{P}_3 \hat{\tilde{D}}_3^* c_{31} \\ \hat{P}_2 \hat{\tilde{D}}_2^* c_{23} = \hat{P}_3 \hat{\tilde{D}}_3^* c_{32}, \end{cases} \tag{2.24}$$

donde el parámetro de tamaño de la población (N) se cancela y $\hat{P}_1$, $\hat{P}_2$ y $\hat{P}_3$ son las proporciones estimadas de los tres grupos de la población.

Pueden aplicarse mínimos cuadrados lineales para resolver el sistema; alternativamente, Heckathorn [18, 41] propuso un enfoque llamado suavizado de datos. La idea básica del suavizado de datos es que si las aristas en la red son recíprocas, si todos los grupos reclutan con la misma eficacia (es decir, para cualquier grupo X, el número de encuestados reclutados por X (RB_X) es igual al número de reclutamientos del grupo X (RO_X), $RB_X = R_{XX} + R_{XY} + \ldots + R_{XN} = R_{XX} + R_{YX} + \ldots + R_{NX} = RO_X$) y si los reclutamientos de redes personales son aleatorios, entonces los reclutamientos (R) de grupos cruzados serán iguales para cada par de grupos, es decir, para cualquier grupo X e Y, $R_{XY} = R_{YX}$.

En el proceso de suavizado de datos, cada elemento R_{XY} se transforma en $c_{XY}\hat{E}_X RB$, donde c_{XY} es la probabilidad de transición de grupos de la muestra, $\hat{E}_X$ es el equilibrio de Markov dada la matriz de probabilidad de transición de grupos muestral C y RB es el número total de reclutamientos en la muestra. El propósito de tal transformación es hacer que la matriz de probabilidad de transición de grupos transformada mantenga las proporciones de selección originales entre grupos e iguale las sumas de filas y columnas. El siguiente paso es utilizar la media de estos recuentos para obtener una matriz de probabilidad de transición de grupos suavizada $\tilde{R}$ de la siguiente manera:

$$\begin{aligned} \tilde{R} &= \begin{bmatrix} c_{11}\hat{E}_1 RB & \frac{c_{12}\hat{E}_1 RB + c_{21}\hat{E}_2 RB}{2} & \ldots & \frac{c_{1M}\hat{E}_1 RB + c_{M1}\hat{E}_M RB}{2} \\ \frac{c_{12}\hat{E}_1 RB + c_{21}\hat{E}_2 RB}{2} & c_{22}\hat{E}_2 RB & \ldots & \frac{c_{2M}\hat{E}_2 RB + c_{M2}\hat{E}_M RB}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{c_{1M}\hat{E}_1 RB + c_{M1}\hat{E}_M RB}{2} & \frac{c_{2M}\hat{E}_2 RB + c_{M2}\hat{E}_M RB}{2} & \ldots & c_{MM}\hat{E}_M RB \end{bmatrix} \\ &= \begin{bmatrix} \tilde{R}_{11} & \tilde{R}_{12} & \ldots & \tilde{R}_{1M} \\ \tilde{R}_{21} & \tilde{R}_{22} & \ldots & \tilde{R}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{R}_{M1} & \tilde{R}_{M2} & \ldots & \tilde{R}_{MM} \end{bmatrix}. \end{aligned} \tag{2.25}$$

Basándonos en $\tilde{R}$, las probabilidades de selección se recalculan en (2.24) y las ecuaciones en exceso que causan el problema de la sobredeterminación se vuelven redundantes. Por ejemplo, en función de las proporciones de selección suavizadas y los grados estimados, las estimaciones de las proporciones de población suavizada de cada grupo se calcula de la siguiente manera en un sistema con M grupos:

$$\begin{cases} 1 = \hat{P}_1 + \hat{P}_2 + \hat{P}_3 + \ldots + \hat{P}_M \\ \hat{P}_1 \hat{\tilde{D}}_1^* \tilde{c}_{12} = \hat{P}_2 \hat{\tilde{D}}_2^* \tilde{c}_{21} \\ \hat{P}_1 \hat{\tilde{D}}_1^* \tilde{c}_{13} = \hat{P}_3 \hat{\tilde{D}}_3^* \tilde{c}_{31} \\ \ldots \\ \hat{P}_1 \hat{\tilde{D}}_1^* \tilde{c}_{1M} = \hat{P}_M \hat{\tilde{D}}_M^* \tilde{c}_{M1}, \end{cases} \tag{2.26}$$

donde $\tilde{c}_{XY} = \tilde{R}_{XY} / \sum_{K \in \{1,\ldots,M\}} \tilde{R}_{XK}$. Es importante notar que el método de suavizado de datos no altera el grado promedio. Esto puede no ser siempre el caso, pero en la práctica ha demostrado ser una suposición admisible [6].

Relación entre RDS I y RDS II

RDS II está estrechamente relacionado con el estimador de RDS I, cuando se consideran variables categóricas, ya que RDS I no es adaptable a la estimación de cantidades continuas.

Tanto el estimador RDS I como RDS II son asintóticamente insesgados [5, 6, 18]. Formalmente, consideramos todos los términos P_A del sistema RDS I de (2.26). Para cualquier grupo X,

$$\begin{aligned} \hat{P}_X &= \frac{\hat{P}_A \hat{\tilde{D}}_A \tilde{c}_{AX}}{\hat{\tilde{D}}_X \tilde{c}_{XA}} \\ &= \frac{\hat{P}_A \hat{\tilde{D}}_A \dfrac{\tilde{R}_{AX}}{\sum_K \tilde{R}_{AK}}}{\hat{\tilde{D}}_X \dfrac{\tilde{R}_{XA}}{\sum_K \tilde{R}_{XK}}}, \end{aligned} \tag{2.27}$$

luego se tiene

$$
\begin{aligned}
\sum_X \hat{P}_X &= \sum_X \frac{\hat{P}_A \hat{\tilde{D}}_A \frac{\tilde{R}_{AX}}{\sum_K \tilde{R}_{AK}}}{\hat{\tilde{D}}_X \frac{\tilde{R}_{XA}}{\sum_K \tilde{R}_{XK}}} \\
&= \frac{\hat{P}_A \hat{\tilde{D}}_A}{\sum_K \tilde{R}_{AK}} \sum_X \frac{\tilde{R}_{AX} \sum_K \tilde{R}_{XK}}{\hat{\tilde{D}}_X \tilde{R}_{XA}} \\
&= \frac{\hat{P}_A \hat{\tilde{D}}_A}{\sum_K \tilde{R}_{AK}} \sum_X \frac{\sum_K \tilde{R}_{XK}}{\hat{\tilde{D}}_X} \\
&= 1.
\end{aligned}
\tag{2.28}
$$

Despreciando los encuestados iniciales "semillas", el número de participantes de tipo X reclutados en el estudio es el mismo que el número de participantes de tipo X en la muestra, es decir, $\sum_K \tilde{R}_{XK} = n_X$.

Si se considera (2.28) para $\hat{P}_A$ se tiene

$$
\begin{aligned}
\hat{P}_A &= \frac{n_A}{\hat{\tilde{D}}_A} \left(\sum_X \frac{n_X}{\hat{\tilde{D}}_X} \right)^{-1} \\
&= \frac{n_A}{\hat{\tilde{D}}_A} \left(\sum_X \frac{n_X}{n_X / \sum_{i \in s \cap X} \delta_i^{-1}} \right)^{-1} \\
&= \frac{n_A}{\hat{\tilde{D}}_A} \left(\sum_X \sum_{i \in s \cap X} \delta_i^{-1} \right)^{-1} \\
&= \frac{n_A}{\hat{\tilde{D}}_A} \frac{1}{\sum_{i \in s} \delta_i^{-1}} \\
&= \frac{n_A}{n_A / \sum_{i \in A \cap s} \delta_i^{-1}} \frac{1}{\sum_{i \in s} \delta_i^{-1}},
\end{aligned}
\tag{2.29}
$$

con un poco de manipulación obtenemos

$$\hat{P}_A = \frac{\sum_{i \in A \cap s} \delta_i^{-1}}{\sum_{i \in s} \delta_i^{-1}}, \tag{2.30}$$

produciendo la forma exacta de (2.8), es decir, el estimador RDS II. Por lo tanto, siempre que se utilice el suavizado de datos, RDS I y RDS II coinciden.

2.3.4. Estimación de la varianza

Método Bootstrap

La precisión de una estimación de la muestra generalmente se mejora al proporcionar un intervalo de confianza (IC), que proporciona un rango dentro del cual se espera encontrar el parámetro poblacional real con cierto nivel de certeza. Debido al complejo diseño de la muestra de RDS, los IC basados en el muestreo aleatorio simple son generalmente más estrechos de lo esperado [18, 42, 43]. En consecuencia, es habitual emplear métodos bootstrap para construir intervalos de confianza en torno a estimaciones de RDS.

Salganik [42] propuso un procedimiento bootstrap ampliamente utilizado en las estimaciones RDS para generar IC. El procedimiento es el siguiente:

1. Se divide a los encuestados de la muestra en dos grupos según la propiedad de sus reclutadores, es decir, aquellos que son reclutados por nodos de tipo A (A_{rec}) y aquellos que son reclutados por nodos de tipo B (B_{rec});

2. Se selecciona aleatoriamente un encuestado de la muestra. Si el encuestado tiene la propiedad A, entonces el siguiente encuestado se elige al azar de A_{rec}; de lo contrario, el siguiente encuestado se elige al azar de B_{rec}. Se continua extrayendo un nuevo encuestado hasta que se alcance el tamaño de muestra original;

3. Se calcula la estimación de RDS basándose en la muestra replicada;

4. Se repiten los pasos (2) y (3), R veces y se calculan las R estimaciones bootstrap;

5. Las estimaciones intermedias del 90 % al 95 % de las R estimaciones bootstrap ordenadas se utilizan luego como el intervalo de confianza estimado.

Estimación de la varianza basada en MCMC

Para tener en cuenta las probabilidades de selección de distribuciones de probabilidad no uniformes y la estructura de Cadena de Markov Monte Carlo (MCMC) de la muestra RDS, Volz y Heckathorn [6] desarrollaron un estimador para la varianza de $\hat{P}_A$ empleando el estimador RDS II con variables categóricas.

Para obtener el estimador $\hat{V}(\hat{P}_A)$, se parte de la estimación de P_A en (2.8) y en que la estimación del grado promedio $\hat{\bar{D}}_U = n/\sum_{i\in s}\delta_i^{-1}$ es constante. Si se escribe $\hat{P}_A$ en función de $\Psi_i = \hat{\bar{D}}_U\delta_i^{-1}I_A(i)$, donde $I_A(i)$ es la función indicadora que toma el valor 1 si $i \in A$ y 0 de lo contrario, se tiene que:

$$\begin{aligned}
\hat{P}_A &= \frac{\sum\limits_{i\in A\cap s}\delta_i^{-1}}{\sum\limits_{i\in s}\delta_i^{-1}} \\
&= \frac{\sum\limits_{i\in A\cap s}\delta_i^{-1}}{n/\hat{\bar{D}}_U} \\
&= \frac{1}{n}\hat{\bar{D}}_U\sum_{i\in A\cap s}\delta_i^{-1} \\
&= \frac{1}{n}\sum_s \hat{\bar{D}}_U\delta_i^{-1}I_A(i) \\
&= \frac{1}{n}\sum_s \Psi_i.
\end{aligned} \tag{2.31}$$

Luego para encontrar la varianza de la expresión anterior se debe encontrar:

$$\begin{aligned}
V(\Psi_1+\cdots+\Psi_n) &= V(\Psi_1+\cdots+\Psi_{n-1}) + V(\Psi_n) + 2cov(\Psi_1+\cdots+\Psi_{n-1},\Psi_n) \\
&= V(\Psi_1+\cdots+\Psi_{n-2}) + V(\Psi_{n-1}) + V(\Psi_n) + 2cov(\Psi_1+\cdots+\Psi_{n-2},\Psi_{n-1}) \\
&\quad +2cov(\Psi_1+\cdots+\Psi_{n-1},\Psi_n) \\
&\quad \vdots \\
&= \sum_s V(\Psi_i) + 2cov\left(\sum_{j<i}\Psi_j,\Psi_i\right).
\end{aligned}$$

Se considera el segundo término de la suma anterior,

$$\begin{aligned} 2cov(\Psi_1 + \cdots + \Psi_{m-1}, \Psi_m) &= E(\Psi_1 + \cdots + \Psi_{m-1} - (m-1)E[\Psi])(\Psi_m - E[\Psi]) \\ &= -(m-1)E[\Psi]^2 + \sum_{i=1}^{m-1} E[\Psi_i \Psi_m]. \end{aligned}$$

En la ecuación anterior, se debe calcular el valor esperado del producto $\Psi_i \Psi_m$. Usando la definición de Ψ_i, se tiene

$$E[\Psi_i \Psi_m] = Pr[i \in A] \cdot E[\Psi_i | i \in A] \cdot Pr[m \in A | i \in A] \cdot E[\Psi_m | m \in A].$$

Donde se estima de la siguiente manera: $Pr[i \in A]$ se estima con n_A/n, $E[\Psi_i | i \in A]$ se estima con $\sum_{s \cap A} \Psi_k / n_A$. La probabilidad que la unidad m también sea de tipo A viene dado por: $Pr[m \in A | i \in A]$ y se estima con $\left(c^{|m-i|}\right)_{AA}$, donde c es el elemento de la matriz de probabilidades de transición estimada de A a A, $|m-i|$ es la distancia entre las unidades de la muestra m e i en la red.

Entonces $E[\Psi_i \Psi_m]$ se estima como:

$$\begin{aligned} E[\Psi_i \Psi_m] &= \frac{n_A}{n} \frac{\sum_{s \cap A} \Psi_k}{n_A} \left(c^{|m-i|}\right)_{AA} \frac{\sum_{s \cap A} \Psi_k}{n_A} \\ &= \hat{P}_A \left(c^{|m-i|}\right)_{AA} \frac{1}{n_A} \sum_{s \cap A} \Psi_k \\ &= \hat{P}_A \left(c^{|m-i|}\right)_{AA} \frac{\hat{\hat{D}}_U}{\hat{\hat{D}}_A} n \\ &= \frac{n}{n_A} \hat{P}_A^2 \left(c^{|m-i|}\right)_{AA}, \end{aligned}$$

donde $\frac{\hat{\hat{D}}_U}{\hat{\hat{D}}_A} = \frac{n}{n_A} \hat{P}_A$.

Luego al reducir la varianza de la suma de Ψ_i a la suma de las varianzas y covarianzas de Ψ_i, se encuentra que:

$$\begin{aligned} V(\Psi_1 + \cdots + \Psi_n) &= \hat{V}\left(n\hat{P}_A\right) \\ &= n^2 \hat{V}\left(\hat{P}_A\right) \\ &= n\hat{V}(\Psi) - \hat{P}_A^2 n(n-1) + \frac{2n\hat{P}_A^2}{n_A} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(c^{|i-j|}\right)_{AA}. \end{aligned}$$

Resolviendo para $\hat{V}\left(\hat{P}_A\right)$ y recordando que $\hat{P}_A = \frac{1}{n} \sum_s \Psi_i$, se tiene

$$\hat{V}\left(\hat{P}_A\right) = \hat{V}_{\hat{P}_A} = \hat{V}_1 + \frac{\hat{P}_A^2}{n}\left((1-n) + \frac{2}{n_A}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(c^{|i-j|}\right)_{AA}\right), \tag{2.32}$$

donde $\hat{V}_1 = \frac{\hat{V}(\Psi_i)}{n} = \frac{1}{n(n-1)}\sum_s (\Psi_i - P_A)^2$ y $|i-j|$ indica la distancia en las olas de la red entre los encuestados i y j.

Desafortunadamente, el estimador de varianza anterior no es insesgado por un par de razones. En primer lugar, $\hat{\tilde{D}}_U$ se incluye en cada término de Ψ_i y, por lo tanto, afectará a la covarianza entre los Ψ_i. Sin embargo, para un tamaño de muestra suficiente, la varianza de $\hat{\tilde{D}}_U$ es generalmente muy pequeña, ya que las probabilidades de selección para esta cantidad son proporcionales a su tamaño. En segundo lugar, las probabilidades de transición c no se conocen en general y por tanto deben estimarse. Aunque la estimación de la varianza no es insesgada, Volz y Heckathorn [6] mostraron empleando estudios de simulación, que en la mayoría de escenarios presenta un rendimiento satisfactorio.

2.3.5. Estimador RDS SS

Gile [44] presenta un estimador que requiere del conocimiento del tamaño de la población. Gile [44] basa su estimación en el muestreo sucesivo[4] y refleja la naturaleza sin reemplazo del proceso de muestreo RDS. El estimador tiene una forma similar al estimador RDS II, sólo que en lugar del peso correspondiente al grado de cada nodo, se incluye una estimación de la probabilidad de inclusión de cada vértice empleando el muestreo sucesivo. El procedimiento básico para calcular el estimador RDS SS es el siguiente:

1. Se inicializa un tamaño de unidad para la función de mapeo de probabilidad de inclusión $f^0(k)$, que asigna a cada unidad k una probabilidad π,

$$f^0(k) = \frac{k}{N}\sum_l \frac{v_l}{l}, \tag{2.33}$$

donde v_l es el número de encuestados con grado l en la muestra. La inicialización asegura que $f^0(k)$ sea proporcional a k.

[4] Sea una población de N unidades, denotadas por los índices $1, \ldots, N$ con unidades de tamaños variables representadas por δ_i. El esquema de muestreo sin reemplazo siguiente, se denomina muestreo sucesivo:

- Se muestrea la primera unidad de la población con probabilidad proporcional al tamaño de la unidad δ_i.
- Se selecciona cada unidad subsiguiente, secuencialmente, de entre las que aún no se han muestreado, con probabilidad proporcional al tamaño en relación con todas las unidades que aún no se han muestreado.

2. Se estima iterativamente la distribución de grados de la población. Para $i = 1, \ldots, r$:

   a) Se estima el número de individuos con grado k en la población:

$$\mathbb{N}_k^i = \frac{N \frac{v_k}{f^{i-1}(k)}}{\sum_l \frac{v_l}{f^{i-1}(l)}}. \tag{2.34}$$

   Este procedimiento utiliza el tamaño de la población N como parámetro conocido.

   b) Se estiman las probabilidades de inclusión para los vértices de la población de $\left\{\mathbb{N}_k^i\right\}$.

   Esto se logra mediante la simulación de M muestras sucesivas de tamaño n de $\left\{\mathbb{N}_k^i\right\}$ y la probabilidad de inclusión de un nodo con grado k se puede estimar mediante

$$f^i(k) \approx \frac{U_k + 1}{M\mathbb{N}_k^i + 1}, \tag{2.35}$$

   donde U_k es el número total de unidades observadas con grado k de las M muestras sucesivas.

3. Después de r iteraciones, $f^r(k)$ se usa como una aproximación de la probabilidad de inclusión para los vértices de grado k, es decir, $Pr(k) \propto f^r(k)$. Sustituyendo δ_i con $f^r(\delta_i)$ en el estimador RDS II (considerando una propiedad binaria tal que cada individuo pertenece al grupo A o al grupo B), la estimación de la proporción poblacional se convierte en:

$$\hat{P}_A = \frac{\sum_{i \in A \cap s} f^r(\delta_i)^{-1}}{\sum_{i \in s} f^r(\delta_i)^{-1}}, \tag{2.36}$$

luego para una característica de tipo continua, la estimación de la media población es:

$$\hat{\bar{y}} = \frac{\sum_{i \in s} f^r(\delta_i)^{-1} y_i}{\sum_{i \in s} f^r(\delta_i)^{-1}}. \tag{2.37}$$

Es recomendable utilizar $M = 2000$ y $r = 3$. El estimador RDS SS ha demostrado un rendimiento superior a otros estimadores en simulaciones con redes de 1000 vértices, con una gran fracción de muestreo (más del 50 % del tamaño de la red) y cuando hay una gran diferencia entre el muestreo con reemplazo y

sin reemplazo. Sin embargo, en una comparación realizada por Tomas y Gile [45] del desempeño de varios estimadores RDS en condiciones de reclutamiento diferencial y falta de respuesta, el estimador RDS SS no superó a otros estimadores en muchas situaciones.

Se debe tener en cuenta que el estimador RDS SS depende del conocimiento del tamaño real de la población, que generalmente no se conoce para las poblaciones ocultas.

2.3.6. Otros enfoques

En una revisión sistemática Abdesselam et al. [46] indican que se han desarrollado 10 estimadores de prevalencia y 2 modificaciones complementarias para un estimador RDS existente (hasta octubre 2019) excluyendo el estimador naive, que es simplemente la media muestral utilizada para la inferencia en el muestreo aleatorio simple (MAS). El desarrollo de los estimadores RDS se puede ver en la tabla 2.1.

Tambien Abdesselam et al. [46] muestran los resultados de la evaluación de los estimadores RDS más comúnmente utilizados y se muestra en la tabla 2.2.

Tabla 2.1: Desarrollo de los estimadores RDS a partir de la introducción de RDS en 1997.

Año	**Autor**	**Estimador RDS**	**Lo más destacado del estimador**
2002	Heckathorn	$RDS\ HK_1^{**}$	La estimación incluye la posestratificación para controlar las diferencias en el tamaño de la red y la agrupación entre grupos.
2004	Salganik and Heckathorn	$RDS\ I^{**}$	Estima el número medio de aristas cruzadas de cada grupo al otro grupo.
2007	Heckathorn	$RDS\ HK_2^{**}$	Es similar a $RDS\ HK_1^{**}$ pero se ajusta para el reclutamiento diferencial y se puede utilizar para estimar variables continuas.
2008	Volz y Heckathorn	$RDS\ II^{**}$	Estima la probabilidad de un individuo en la población por su tamaño de red y supone que las probabilidades de muestreo son proporcionales al grado.
2011	Gile	$RDS\ SS^{**}$	El estimador $RDS\ SS^{**}$ tiene una forma similar al estimador $RDS\ II$ con la excepción que toma en consideración el muestreo sin reemplazo y requiere que se conozca el tamaño de la población.
2013/2016	Lu et al. y Malmros et al.	$RDS\ DN$	Ajusta $RDS\ I$ y $RDS\ II$ por el grado de entrada y salida, que ya no se basa en la suposición que la red no está dirigida.
2013	Lu	RDS^{IEGO}	Incorpora datos de la red de ego en $RDS\ I$ y $RDS\ II$, pero demuestra que $RDS\ I$ + red del ego(RDS^{IEGO}) es superior.
2015	Gile y Handcock	$RDS\ MA^{*}$	Se basa en el modelo gráfico aleatorio de familias exponenciales para realizar estimaciones de la prevalencia de la población y se supone que se ajusta a la selección de semillas y al reclutamiento diferencial.
2015	Aronow y Crawford	$RDS\ II^{*}$ no paramétrico	La estimación se basa en supuestos menos estrictos; sin embargo, las condiciones enumeradas no se pueden comprobar.
2016	Selvaraj et al.	$RDS\ MOD$	La estimación tiene un enfoque similar al de $RDS\ II$. La modificación es el multiplicador constante adicional que incorpora los pesos de la muestra y la información sobre la correlación entre las unidades de la muestra, capturando el efecto de agrupamiento.
2017	Berchenko et al.	RDS^{*} Modelo de proceso de conteo	Este estimador de máxima verosimilitud requiere los conteos de individuos y el tiempo de entrevista de los participantes. Este estimador descarta el modelo de camino aleatorio homogéneo para un modelo de epidemia estocástica.

* No se basa en la teoría de primer orden de Markov.

** Estimadores más utilizados en la literatura.

Fuente: Abdesselam et al. [46]

Tabla 2.2: Evaluación de las violaciones en las condiciones de los estimadores RDS más comúnmente utilizados.

Estimador	Selección de semillas	Reclutamiento diferencial	Efectividad y no respuesta en el reclutamiento	Relación no recíproca	Muestreo sin reemplazo	Falta de información de grados	Observación adicional
Naive	Muestra un gran sesgo [31]	Muestra un gran sesgo [31, 47]. No se ve afectado por diferentes profundidades de reclutamiento [48]	Muestra un gran sesgo [31, 49]	No afectado [50]	No afectado [50]	No se ve afectado porque esta estimación no utiliza información sobre grados [50]	▪ Ignora el diseño de muestra de RDS [5]. ▪ El estimador naive supera a *RDS I* y *RDS II* cuando el marco de muestreo está disponible [51] y cuando la red de la población está distribuida normalmente [50].
RDS HK_1	Susceptible a un sesgo sustancial de la selección de semillas [47]	Rendimiento mejor que la estimación naive en diversas condiciones, excepto en el modelo aleatorio [47]	Sesgo significativo introducido cuando el reclutamiento no se completa y no alcanza el número de oleadas calculado [47]	No evaluado	Si la fracción de muestreo es muy pequeña, la estimación no se ve afectada [6]	No evaluado	▪ Cuando se tiene en cuenta la transmisión de enfermedades y la distribución de la red, en general este estimador supera al estimador naive [47].
RDS I	Susceptible al sesgo de semillas [48]. Sin embargo, cuando las semillas seleccionadas son muy poco representativas de la población, superan a *RDS II*, *RDS SS* y *RDS* HK_2 [45]	Supera a *RDS II* *RDS* HK_2 y *RDS SS* con la ausencia de actividad diferencial y una gran fracción de muestreo [45]. Puede ocurrir un alto sesgo cuando se produce un muestreo no uniforme [52]	Controla el sesgo en fracciones de muestreo bajas [31] y en profundidades de reclutamiento altas [48]	Susceptible de sesgo en la red dirigida	Sesgo inducido a probabilidades de muestreo desigual entre diferentes distribuciones de red [52]	Sesgo modesto, se observa un sesgo mayor en el modelo ER y BA [53]	▪ Susceptible a cuellos de botella en cualquier parte de la red [43]. ▪ Muy sensible al número desproporcionado de contrataciones de un grupo a otro grupo - fuertemente sesgado a subestimar la prevalencia de la población [45]. ▪ Estimación fuertemente asociada a la distribución de la red [53].
RDS HK_2	Parece ajustarse al sesgo introducido por la selección de semillas [31]	La actividad diferencial conduce a un sesgo significativo [31]	Controles de sesgo en fracciones de muestreo bajas [31]	No evaluado	Sesgo insignificante para fracciones de muestreo pequeñas ($\leq$ 20 %) y sesgo bajo por debajo del 40 % para varias distribuciones de red [54]	Sesgo susceptible de moderado a bajo [54]	▪ Funciona de manera similar a *RDS I* en diversas condiciones [45]. ▪ El estimador general funciona bien cuando la fracción de muestreo está por debajo del 20 % y hay una homofilia de baja a moderada ($< 0{,}7$) [54].
RDS II	Sesgo inducido por la selección de semillas [31]. Superar a *RDS I* [48]. *RDS II* se ajusta para la selección de semillas [50]	La actividad diferencial conduce a un sesgo significativo [31, 55]	Sesgo introducido cuando el tamaño de la muestra excede el 10 % de la población objetivo [55] y a profundidades de reclutamiento bajas [48]	Susceptible de sesgo en la red dirigida; pero se puede ajustar [55]	No es un problema con el tamaño de la muestra que va de 500 a 1000 [55]	Sesgo introducido cuando el tamaño de la muestra supera el 10 % de la población objetivo [48]. Se observa un mayor sesgo en los modelos ER y BA [53]	▪ Susceptible a cuellos de botella en cualquier parte de la red [43]. ▪ Susceptible a una alta distribución de homofilia [53]. ▪ Una fracción de muestreo alta conduce a un sesgo sustancial [31]. ▪ La precisión de *RDS II* en varias distribuciones de red fue mejor que la de *RDS* HK_2 [45]. ▪ Estimaciones fuertemente asociadas con el grado de la red.
RDS SS	Susceptible al sesgo de semillas a baja profundidad de reclutamiento [48]	Sesgo cuando hay homofilia [45, 48]	Sesgo a profundidades de reclutamiento bajas [48]	No evaluado	No aplica ya que la estimación considera muestras sin reemplazo [44]	Susceptible de sesgo cuando el grado informado es incorrecto [44]	▪ Se requiere el tamaño de la población. ▪ En general, no existe diferencias significativas [48] entre *RDS I* y *RDS II*. ▪ Existe poca diferencia entre *RDS II* y *RDS SS* [48].

Fuente: Abdesselam et al. [46]

Con relación a la tabla 2.2 Abdesselam et al. [46] indican que la mayoría de los estudios, incluidos Gile et al. [35] demostraron que un estimador puede funcionar mejor bajo ciertas condiciones, pero en general, ningún estimador supera en todos los escenarios a otro. El estimador $RDS\ II$ es el estimador más utilizado y funciona muy bien cuando se utiliza una pequeña fracción de muestreo. Verdery et al. [56] evaluaron la mayoría de los estimadores, tanto en condiciones de RDS ideales como con datos reales y al igual que otros, concluyeron que los estimadores más comunes utilizados funcionan de manera similar.

Capítulo 3

Encuesta RDS de minorías étnicas

3.1. Introducción

Obtener datos cuantitativos fiables sobre los grupos minoritarios en general y las minorías étnicas (ME) en particular, es un gran desafío para los profesionales de la investigación de encuestas [64]. Debido al pequeño tamaño de su población general, las dificultades para acceder a ellos y sus tasas de respuesta más bajas [65, 66], obtener una buena representación de estos grupos es muy costoso y a menudo se subestiman en las encuestas oficiales [67].

Se han desarrollado diferentes enfoques para optimizar el costo, consistencia y precisión de las técnicas de muestreo basadas en la probabilidad. Por ejemplo, el sobremuestreo en áreas con mayores proporciones de residentes de minorías étnicas, o pidiendo a los encuestados de ME que indiquen cuáles de sus conocidos pertenecen al mismo grupo de ME (enumeración focalizada), realizando un muestreo de los vecinos de los encuestados de ME (muestreo adaptativo por conglomerados), ejecutando un muestreo por apellidos (para minorías étnicas con antecedentes migratorios) o muestreo en los lugares donde se reúnen estas poblaciones (conocido como TLC) [67, 68].

Si bien el muestreo probabilístico[1] permite producir estimaciones no sesgadas de parámetros poblacionales con datos de encuestas, su implementación requiere la movilización de muchos recursos económicos y humanos, algo que está fuera del alcance de muchos presupuestos de investigación. Por otro lado, estas

[1] Aplicado a un universo, donde cada unidad tienen una probabilidad (superior a cero) de ser seleccionado en la muestra y esta probabilidad se puede determinar con precisión.

técnicas pueden no ser factibles cuando no se dispone de un marco muestral para la población o pueden resultar ineficientes cuando el foco de interés son subgrupos específicos de ellas [69]. Además, algunas de estas técnicas, en particular las que se basan en supuestos sobre la distribución geográfica de las poblaciones de ME, pueden estar sujetas a sesgos [4, 67, 70]. Los enfoques de muestreo basados en la ubicación tienden a ignorar a los individuos que no encajan en el modelo (es decir, viven en vecindarios mixtos o blancos no hispanos según su composición étnica, no van a la mezquita ni usan cabinas telefónicas, etc.), ofreciendo así una descripción sesgada del grupo ME, que suele ser más heterogéneo de lo que muestra la descripción [67, 71, 72]. Por ejemplo, varios estudios en el Reino Unido han demostrado que el uso de muestreo estratificado, con afijación proporcional según el tamaño de la población de ME en el área, da como resultado un sesgo urbano significativo para algunas minorías étnicas y la subrepresentación de otras, que son geográficamente más dispersos, como las comunidades chinas o las judías [70]. Por último, la mayoría de estas técnicas se han diseñado y optimizado para su uso con los métodos de recopilación de datos administrados por el encuestador, como las entrevistas cara a cara o por teléfono y no son factibles cuando se utilizan métodos de encuesta web, porque para la mayoría de las poblaciones objetivo este modo de recolección de datos carece de un marco de muestreo válido [73].

Por todas las razones anteriores, el uso de métodos de muestreo no probabilísticos se está multiplicando en diferentes campos de investigación en un contexto en el que aumentan las preocupaciones sobre la cobertura y la falta de respuesta, junto con los crecientes costos de los métodos tradicionales de investigación de encuestas [74]. Entre estos diseños no probabilísticos se encuentran las técnicas de muestreo en red, también conocidas como métodos de referencia en cadena, como el muestreo en bola de nieve y muestreo dirigido por los participantes. Una clara ventaja de los métodos de referencia en cadena sobre el muestreo probabilístico es que los límites y la estructura de la población se definen de adentro hacia afuera, un enfoque ascendente que ha resultado útil para tratar con poblaciones minoritarias, de difícil acceso y/o ocultas [75].

En la siguiente sección, comenzaremos mostrando la necesidad de datos sobre minorías étnicas, en particular describimos algunos estudios RDS desarrollados recientemente en estas poblaciones. En la sección 3.3, se presenta una aportación original que considera el problema de encuestar mediante la metodología RDS a unos grupos étnicos en Ecuador.

3.2. Necesidad de datos sobre minorías étnicas

Dentro de la investigación étnica, los científicos de las Ciencias de la Salud quieren descubrir las causas y los procesos de las enfermedades, mientras que los encargados de formular políticas y planificar la salud quieren satisfacer las necesidades de los grupos étnicos minoritarios (la falta de datos puede obstaculizar estos objetivos). El análisis histórico revela motivos como el deseo de revertir las desventajas sociales y de salud de los grupos étnicos minoritarios [76]. De hecho, debido a una mezcla compleja de razones genéticas, ambientales y de comportamiento, los grupos minoritarios étnicos a menudo corren un mayor riesgo de resultados adversos para la salud que otros segmentos de la población. Estos grupos de la población han sido un foco de planificación y establecimiento de objetivos de salud durante décadas. El cumplimiento de muchos de estos objetivos de planificación política, requiere que existan estrategias de diseño estadístico adecuadas para muestrear las minorías y, por lo tanto, recopilar datos en estudios epidemiológicos y otros estudios de salud dirigidos a la población en los que se evalúa el riesgo. Más allá de la experiencia de salud distintiva de los principales grupos minoritarios (por ejemplo, afroamericanos e hispanos) está la diversidad que existe dentro del grupo. Esta diversidad ha aumentado aún más la demanda y, por lo tanto, la dificultad de obtener datos adecuados para estos subgrupos minoritarios aún más pequeños [77].

3.2.1. Estudios RDS en poblaciones de minorías étnicas

Se han desarrollado estudios RDS de personas en riesgo de contraer el VIH y usuarios de drogas inyectables, la comunidad LGBTI, niños de la calle, etc. Sin embargo, en estas investigaciones no se estudian las ME como centro de la investigación, más bien en algunos casos son una característica más del análisis. De nuestro conocimiento se han desarrollado pocos estudios que aplican la metodología RDS para el análisis de poblaciones de minorías étnicas, entre estos estudios encontramos los siguientes:

- "Eficacia del muestreo dirigido por los participantes para la evaluación de la salud de las poblaciones minoritarias". En este trabajo, se realiza una encuesta sobre comunicación sanitaria en la población general de Guam utilizando el método RDS para generar estimaciones de población razonables de las poblaciones minoritarias y generales [57].

- "Aumento de la participación de las minorías étnicas en los ensayos clínicos de abuso de sustancias:

lecciones aprendidas en la Red de ensayos clínicos del Instituto Nacional sobre el Abuso de Drogas". En este documento se incluyen tres estrategias prometedoras para mejorar la inclusión de minorías étnicas en la Red de Ensayos Clínicos, entre estos el muestreo dirigido por los participantes [58].

- "Evaluación del muestreo dirigido por los participantes para estudios de redes en contextos etnográficos". En este artículo, se emplea RDS como parte de un proyecto de investigación de redes sociales, donde se reclutan 330 residentes mayores de 17 años de una comunidad predominantemente aborigen (92 %) en Nain, en el norte de Labrador, Canadá, para entrevistas en redes sociales sobre el intercambio de alimentos, vivienda, salud pública y tradiciones comunitarias [59].

- "Desenmascarar los determinantes y resultados de la salud de las Primeras Naciones urbanas mediante un muestreo dirigido por los participantes". El objetivo principal es trabajar en asociación con las partes interesadas aborígenes para generar un conjunto de datos de salud de referencia representativos y culturalmente relevantes para tres comunidades aborígenes urbanas en Ontario, Canadá [60].

3.3. Encuesta RDS de minorías étnicas en Ecuador

RDS se ha utilizado en todo el mundo en cientos de encuestas para investigaciones relacionadas con la salud para establecer políticas sanitarias, como estrategia de muestreo que puede producir estimaciones rigurosas y representativas de prevalencia de enfermedades. RDS se ha utilizado para encuestas sobre evaluación de salud, ensayos clínicos de abuso de sustancias y características sociodemográficas de poblaciones de minorías étnicas [57, 58, 59, 60]. Los resultados de estas encuestas proporcionan datos muy necesarios para asignar fondos, evaluar el éxito de un programa y planificar cualquier intervención y programas de prevención necesarios. En este sentido se presenta a continuación, una de las principales aportaciones de este trabajo: una encuesta realizada mediante el muestreo dirigido por los participantes, para estudiar las condiciones de vida y aspectos socioeconómicos de tres grupos étnicos en el cantón Riobamba, Ecuador. Este trabajo ha sido publicado recientemente en la revista Sustainability (Mullo, H.; Sánchez-Borrego, I.; Pasadas-del-Amo, S. Respondent-Driven Sampling for Surveying Ethnic Minorities in Ecuador. *Sustainability.* **2020**, *12(21)*, 9102, doi.org/10.3390/su12219102.). Factor de impacto (2019): 2.576. Posición 120/265 (Q2) en el listado JCR Environmental Sciences.

3.3.1. Introducción

Como cualquier otro país de América del Sur, Ecuador es étnicamente diverso [78, 79, 80]. La mayor parte de su población se identifica como mestiza (71,93 %), que comprende una herencia mixta amerindia y española y las siguientes minorías: Montubio (7,39 %), Afroecuatorianos (7,19 %), Indígena (7,03 %), Blancos (6,09 %) y otros (0,37 %) [81]. La pobreza y las necesidades básicas insatisfechas afectan a más hogares indígenas y afroecuatorianos que a los de origen mixto [82]. Además, la incidencia de la pobreza extrema en los hogares indígenas y afroecuatorianos es incluso mayor que en las familias mestizas, lo que se traduce en menor acceso a la educación y dificultades para conseguir una vivienda digna [82]. Varios estudios han destacado que la vivienda es uno de los componentes más importantes del bienestar subjetivo y la satisfacción general de una persona [83, 84, 85]. Chica et al. [86] recientemente estudiaron la calidad de vida de los hogares en Colombia y, Chica y Cano [87] estudiaron los precios de las casas utilizando un método de regresión-kriging. Existe evidencia que los problemas con la vivienda y el nivel de vida de algunos grupos de la población y su exclusión social pueden generar problemas de salud [88, 89].

Algunos estudios y datos recopilados de agencias oficiales de estadística destacan la mayor tasa de desempleo y el escaso acceso a la salud, la educación y la vivienda de los pueblos indígenas y otras minorías étnicas en comparación con los de origen mixto [90, 91, 92, 93, 94]. No obstante, la mayoría de estos estudios se basan en fuentes de datos e información sobre las condiciones de vida de las minorías étnicas en el Ecuador como el Censo de población y vivienda (CPV) 2010, la Encuesta de condiciones de vida (ECV) 2014, o la Encuesta Nacional de empleo, desempleo y subempleo (ENEMDU) 2018, pero no son representativos a nivel cantonal y están desactualizados o no toman en cuenta subgrupos específicos de la población indígena (es decir, jóvenes de ME a nivel provincial o cantonal). Recientemente, Rotondi et al. [10] mostraron en un estudio de muestreo dirigido por los participantes realizado en Canadá, que la población indígena en Toronto también está subrepresentada, ya que el Censo de Canada (mediante la Agencia Nacional de Estadística de Canadá) subestima esta población. Se proporciona más información sobre este grupo y sus problemas de salud en [95, 96]. Generalmente, obtener información confiable sobre grupos minoritarios y minorías étnicas, en particular, es un desafío para los profesionales de la investigación de encuestas [11]. Estos grupos generalmente representan un tamaño pequeño de la población general, lo que dificulta el acceso a ellos y generalmente obtienen tasas de respuesta más bajas [97]. Además, hay indicios que los pueblos indígenas pueden estar subrepresentados en la Agencia

Nacional de Estadísticas del Ecuador. Según la Confederación Nacional de Organizaciones Campesinas, Indígenas y Negras (FENOCIN) hasta el 70 % de la población ecuatoriana pertenece a estos grupos sociales [98], pero se resisten a ser identificados como miembros de estos grupos por prejuicios y estigmas sociales [98, 99, 100, 101]. Por lo tanto, son difíciles de alcanzar, a menudo se subestiman en las encuestas oficiales [67] y usar el muestreo tradicional para encuestarlas puede resultar sumamente costoso.

Usar la metodología RDS para aborda el problema de encuestar a una población de indígenas y otras minorías étnicas en el Cantón Riobamba, Ecuador, puede llenar un vacío ya que no se ha hecho tal enfoque para estudiar a este grupo étnico en Ecuador. Estudiamos sus condiciones de vida centrándonos en variables como trabajo, ingresos, vivienda, exclusión social, pobreza, percepción del nivel de vida, etc. Se obtienen estimaciones precisas sobre las condiciones sociales y de vida de estos colectivos utilizando los estimadores más relevantes en RDS. Ilustramos y discutimos algunas características importantes de una encuesta RDS, como los puntajes de homofilia y las gráficas de convergencia y cuello de botella, que normalmente se utilizan para evaluar que se cumplen algunos de los supuestos de esta metodología.

3.3.2. Materiales y métodos

Se desarrolla una encuesta transversal realizada por la Universidad de Granada y la Escuela Superior Politécnica de Chimborazo (ESPOCH) del Ecuador, en colaboración con la FENOCIN. Los criterios de inclusión de los participantes en la encuesta fueron: autoidentificación como indígena, monubio o afro-ecuatoriano, tener entre 18 y 29 años; residir en la ciudad de Riobamba, tener un número de identificación vigente y dar su consentimiento para participar en el estudio. En primer lugar, se evaluó la idoneidad de RDS para encuestar a este grupo de personas. La selección no aleatoria de los encuestados iniciales, denominada semillas, es fundamental, ya que deben tener una gran red social. Fueron seleccionados mediante entrevistas presenciales y telefónicas entre 40 jóvenes líderes que trabajan en temas de interculturalidad, justicia y solidaridad para la FENOCIN. Fueron entrevistados varias veces para asegurar su idoneidad. Se seleccionaron diez semillas, diversas en términos de sexo, edad, estado civil, etnia e instrucción (ver la tabla 3.1). La selección se basó en dos criterios: características personales y número de conexiones dentro de su grupo social.

La metodología RDS requiere que el grupo de interés sea una población oculta y que formen una red social bien conectada.

Tabla 3.1: Muestra inicial (semillas) para la encuesta de muestreo dirigido por los participantes según características sociodemográficas.

Semilla	Sexo	Edad	Estado civil	Grupo Étnico	Nivel de Educación
1	Mujer	29	Casado o unión libre	Indígena	Postgrado
2	Mujer	21	Casado o unión libre	Indígena	Escuela secundaria
3	Mujer	29	Casado o unión libre	Indígena	Universidad
4	Mujer	28	Soltero	Indígena	Universidad
5	Mujer	21	Casado o unión libre	Indígena	Escuela secundaria
6	Mujer	28	Casado o unión libre	Montubio	Postgrado
7	Hombre	26	Casado o unión libre	Indígena	Secundaria
8	Hombre	21	Soltero	Indígena	Universidad
9	Mujer	21	Soltero	Afroecuatoriano	Secundaria
10	Hombre	22	Soltero	Indígena	Escuela secundaria

Las poblaciones indígenas, monubias y afroecuatorianas se han estudiado hasta ahora en las encuestas CPV, ECV y ENEMDU, pero no hay una encuesta disponible centrada en poblaciones étnicas jóvenes excluidas en Ecuador. Más importante aún, a los jóvenes indígenas, monubios y afroecuatorianos les resulta difícil identificarse a sí mismos [98, 99, 100, 101]. Por lo tanto, carecemos de un marco de muestreo confiable para este grupo, lo que dificulta la implementación del muestreo tradicional. Dado que RDS reduce las preocupaciones por la privacidad, puede ser un método conveniente para encuestar a dichas poblaciones [4].

El cantón Riobamba es el hogar de la mayor proporción de jóvenes indígenas en el Ecuador y cuenta con la presencia de jóvenes monubios y afroecuatorianos [102]. La evidencia de diferentes estudios muestra que la población indígena en la provincia de Chimborazo y particularmente en la ciudad de Riobamba, concentra la gran mayoría de la población indígena más pobre del país, con condiciones de vida esenciales lejos de las ideales [92, 93, 94]. La ciudad de Riobamba tiene aproximadamente 39.000 estudiantes en universidades, institutos tecnológicos, etc. La mayoría de los monubios y afroecuatorianos que viven en Riobamba son estudiantes que han emigrado de áreas rurales (de la costa ecuatoriana principalmente) a la ciudad de Riobamba en busca de educación y mejores condiciones de vida. Esto ha llevado a una asociación de diferentes grupos sociales, que se han fusionado en una sola red social [101]. Por tanto, los jóvenes indígenas, monubios y afroecuatorianos son una población oculta, son un grupo socioeconómico interesante y forman una red social, por lo que se encuesta a esta etnia joven mediante RDS.

El cuestionario incluye ocho secciones diferentes que cubre la siguiente información: datos de contacto y elegibilidad, consentimiento informado, sociodemografía, vivienda y hogar, salud, hábitos, prácticas y

uso del tiempo, pobreza, discriminación y satisfacción general con la vida (ver la tabla 3.2). Una vez que los participantes completan la encuesta, se convirten en reclutadores y acceden a un formulario diferente para reclutar nuevos encuestados. Tanto el cuestionario como los formularios de contratación se alojaron en el sitio web (`www.ugremina.com`). Después de recopilar la información de contacto de los nuevos participantes en cada ola, el sistema informático envia un correo electrónico con instrucciones sobre cómo usar el sitio web para completar la encuesta y cómo reclutar hasta tres nuevos participantes. El sistema informático espera su respuesta en las dos semanas siguientes con hasta cuatro textos recordándoles completar el cuestionario e invitar a nuevos compañeros. Cada reclutado recibe un nombre de usuario y una contraseña por correo electrónico para iniciar sesión en el sitio web. Los identificadores tanto del reclutado como de su reclutador se almacenan en una base de datos, lo que permite rastrear las cadenas creadas a partir de cada semilla. Se monitorea el proceso de reclutamiento, asegurándonos que la muestra RDS sea lo suficientemente grande para superar el sesgo potencial introducido con la selección inicial de semillas. Utilizamos los gráficos de convergencia y cuello de botella (Figuras 3.3 y 3.4) para comprobar la evolución de las cadenas de RDS. La muestra final de RDS es consistente con las encuestas CPV 2010 y ENEMDU 2018. Se utiliza un sistema dual de incentivos para promover la contratación, como se suele hacer con las encuestas RDS [4]. El incentivo es el derecho a participar en una lotería donde el premio es un viaje de vacaciones a Galápagos. Los participantes recibien un boleto de rifa inmediatamente después de completar la encuesta web y otro por cada compañero (hasta tres) reclutado con éxito.

Para tener en cuenta los supuestos de RDS, consideramos las gráficas de convergencia y cuello de botella y la razón de homofilia de las variables. La homofilia es la tendencia a asociarse con personas con características similares. Los puntajes de homofilia de la encuesta RDS se muestran en la tabla 3.3 y se interpretan en la siguiente sección. Los gráficos de convergencia muestran el parámetro de población real con el número de reclutados en el eje horizontal. Esta gráfica puede ayudar a evaluar si la muestra está sesgada por el conjunto inicial de semillas. Las parcelas de cuello de botella pueden mostrar diferencias entre semillas. En la siguiente sección se ofrecen e interpretan estas dos gráficas para los datos de la encuesta étnica RDS.

Utilizamos los estimadores más habituales en RDS, que son el estimador de razón RDS I, el estimador RDS II y la versión de Gile y Hanckock [44] para muestreo con reemplazo (RDS SS). Los estimadores RDS SS y RDS II tienen propiedades estadísticas deseables ya que son consistentes y asintóticamente insesgados. Utilizamos el entorno de software R, en particular, la librería RDS [103] y la biblioteca igraph

para dibujar redes sociales.

Tabla 3.2: Variables y redacción de las preguntas de la encuesta RDS de jóvenes urbanos indígenas, monubios y afroecuatorianos del cantón Riobamba, Ecuador.

Componentes	**Variables**	**Preguntas**
Características sociodemográficas	Sexo *	Género
	Edad	¿Cuántos años tienes?
	Estado civil	¿Cuál es tu estado civil?
	Autoidentificación étnica	¿Cómo te identificas según tu cultura y costumbres?
	Cantón *	¿En qué cantón naciste?
	Ropa *	¿Utiliza la ropa tradicional de su grupo étnico?
	Instrucción *	¿Cuál es el nivel más alto de educación que ha completado con éxito?
	Instrucción de la madre	¿Cuál es el nivel más alto de educación que su madre completó con éxito?
	Instrucción del padre	¿Cuál es el nivel más alto de educación que completó con éxito su padre?
Vivienda y hogar	Número de cuartos	¿Cuántos dormitorios hay en su casa o apartamento?
	Número de personas	¿Cuántas personas viven en su hogar?
	Servicio de agua	¿Cuál es la principal fuente de agua en su hogar?
	Servicio de energía	¿Cuál es la principal fuente de iluminación para tu hogar?
Salud	Discapacidad	¿Algún miembro de su hogar tiene alguna discapacidad física, sensorial o mental permanente?
	Visita	¿Fue visitado por la Misión Manuela Espejo?
Hábitos, prácticas y uso del tiempo	Idioma *	¿Qué idioma o idiomas habla comúnmente?
	Idioma de los padres	¿Qué idioma o idiomas usan tus padres contigo?
	Inscripción	Este año escolar, ¿estabas inscrito en...?
	Razón para no inscribirse	Si no se inscribió, ¿cuál fue su principal razón?
	Idioma de la clase	¿Cuál es el idioma principal de las clases en el establecimiento donde estás matriculado?
	Tipo de escuela	¿Qué tipo es el establecimiento donde te inscribiste o registraste?
Pobreza	Trabajo *	¿Qué hiciste la última semana?
	Seguridad social	Si trabajó, ¿qué tipo de seguridad social tiene?
	Ocupación	Si trabajó, ¿en cuál de los siguientes grupos de ocupación trabajó?
	Relación laboral	Si trabajó, ¿en cuál de las siguientes categorías de ocupación trabajó?
	Salario	Si trabajó, ¿cuál fue el pago o el salario mensual total (en dólares y antes de los descuentos) que recibió por trabajar el mes pasado?
	Tierras	¿Usted o algún otro miembro de su hogar tiene tierras destinadas a usos agrícolas (lotes, parcelas o granjas)?
	Cosecha	Durante los últimos 12 meses, ¿cosechaste o recibiste algún producto agrícola de esta tierra?
	Animales	Durante los últimos 12 meses, ¿usted o algún otro miembro de su hogar crió animales, como pollos, pavos, conejillos de indias, conejos, cerdos, ovejas, ganado, etc., en esta granja o tierra?
	Pobreza *	¿Cómo considera su hogar según su condición económica?
Percepción de discriminación	Víctima *	¿Ha sido víctima de episodios de discriminación?
	Satisfacción general	En una escala del 1 al 10, donde 1 significa totalmente infeliz y 10 significa totalmente feliz, ¿cómo es tu satisfacción general considerando todos los aspectos de tu vida?

* Respuestas binarias no condicionadas a respuestas a otras preguntas de la encuesta.

Se ha utilizado una estimación de N = 11,920 (según el CPV 2010). N es necesario para calcular el estimador RDS SS.

Tabla 3.3: Estimaciones de homofilia de reclutamiento para los participantes en la encuesta RDS de jóvenes étnicos urbanos.

Variable	**Estimación homofilia**
Ocupación	1.15
Cosecha	1.09
Número de personas	1.08
Número de cuartos	1.06
Tierra	1.03
Ropa	1.03
Autoidentificación étnica	1.03
Animales	1.02
Estado civil	1.02
Seguridad social	1.02
Instrucción	1.02
Instrucción de la madre	1.02
Servicio de agua	1.01
Instrucción del padre	1.01
Pobreza	1.00
Servicio de energía	1.00
Discapacidad	1.00
Tipo de escuela	1.00
Idioma de la clase	1.00
Satisfacción general	0,99
Víctima	0,99
Idioma	0.98
Trabajo	0.98
Idioma de los padres	0.98
Razón no inscripción	0.96
Inscripción	0.93
Salario	0.93
Sexo	0.93
Cantón	0.93
Edad	0.92
Relación laboral	0,75

3.3.3. Resultados

Como se mencionó anteriormente, se seleccionaron diez semillas iniciales para participar y reclutar hasta tres encuestados más. Cada nuevo encuestado tuvo la oportunidad de reclutar a otros tres nuevos participantes en el estudio. Treinta y dos de los reclutados utilizaron los tres cupones para reclutar, 300 solo dos de ellos, 108 uno y 60 no reclutaron a nadie. Tres de las semillas tuvieron mucho éxito (86 o más reclutadas dentro de sus cadenas), cuatro tuvieron un éxito moderado (de 66 a 85 reclutadas dentro de sus cadenas) y tres tuvieron un desempeño menor (65 o menos reclutadas). La encuesta alcanzó seis oleadas para las 10 semillas (ver Figuras 3.1 y 3.2).

Se enviaron un total de 1510 invitaciones a posibles participantes elegibles y 814 completaron el cuestionario, lo que arroja una tasa de cooperación global del 53,9 %. Los casos válidos y las tasas de cooperación se distribuyen desde la primera ola hasta la sexta, como se muestra en la Figura 3.2.

Condiciones de vida del grupo étnico

Estudiamos las condiciones de vida de este grupo étnico y comparamos las estimaciones de RDS con los valores obtenidos con las encuestas oficiales CPV 2010 y ENEMDU 2018 para los ecuatorianos regulares (valores de color azul) y los pertenecientes a minorías étnicas (valores de color verde). Las estimaciones de RDS y los intervalos de confianza se informan en la tabla 3.4. Calculamos los tres estimadores habituales de RDS (dados en la Sección 3.3.2) para cada característica en estudio y obtuvimos resultados similares, que se informan en la tabla 1 del Apéndice .1.

RDS permite reclutar participantes que normalmente no serían parte de una muestra probabilística en el contexto del estudio de poblaciones ocultas. La edad, el estado civil y las características salariales (para la encuesta ENEMDU de minorías étnicas) caen fuera del intervalo de confianza al 95 % RDS, lo que indica que las personas que se resisten a ser identificadas como parte de esas minorías étnicas (que fueron capturadas por RDS) tienden a ser más jóvenes (21,81 años) que aquellos que no tienen ningún problema con su autoidentificación étnica (23,25 años). Del mismo modo, tienden a ser solteros (81,31 % frente al 55,09 %) y con ingresos medios más bajos ($295,50 frente a $379,64). En contraste, las características sexo, instrucción, idioma, trabajo, seguridad social y pobreza extrema caen dentro del intervalo, lo que demuestra que no existe diferencia para estas variables entre quienes se autoidentifican como parte de la minoría étnica y quienes no lo hacen.

Comparamos las estimaciones de la encuesta Total ENEMDU con los intervalos de confianza RDS al 95 % con la intención de identificar brechas. La tabla 3.4 muestra grandes diferencias en las características socioeconómicas, como el salario para "Total ENEMDU" y la "estimación RDS" ($523,58 en comparación con $295,496), quedando el primero muy por fuera del intervalo de confianza RDS al 95 %. También hay diferencias en la cobertura de instrucción y seguridad social entre estos dos grupos, con los valores totales de ENEMDU fuera de los intervalos de confianza. Además, el 90,63 % de los jóvenes étnicos afirma haber sido ocasionalmente víctima de discriminación.

Siguiendo los mismos argumentos, existen diferencias en la mayoría de las características de la vivienda consideradas en la encuesta (número de personas que viven en una casa, agua y servicio de energía). A pesar del importante esfuerzo que están realizando las administraciones ecuatorianas para evitar la exclusión social y la discriminación de estos grupos étnicos, existen evidencias de diferencias socioeconómicas.

Evaluación de la encuesta RDS

Calculamos las proporciones de homofilia de las características en estudio, como se muestra en la tabla 3.3. La homofilia se calcula como la relación entre el número de reclutados con la misma característica que su reclutador y el número esperado por azar [21]. Un valor de 1 significa que no hay reclutamiento preferencial, mientras que los valores superiores a 1 indican homofilia y los valores inferiores a 1 heterofilia (es decir, un valor un poco superior a 1 indica una homofilia moderada).

Para evaluar el reclutamiento, calculamos las puntuaciones de homofilia para cada variable en estudio. Existen variables socioeconómicas que están más conectadas y presentan homofilia (tendencia a asociarse con aquellas con características similares). El grupo de ocupación en el que trabajan parece tener una homofilia modesta. Para otras características socioeconómicas, como la vestimenta o la seguridad social, la homofilia es muy pequeña. Generalmente, la mayoría de los valores se acercan a 1, lo que indica una homofilia modesta o una heterofilia modesta y, por tanto, un reclutamiento satisfactorio.

Las gráficas de convergencia en la Figura 3.3 indican los valores de la muestra que convergen al parámetro de población real para las variables sexo, nivel de instrucción, estado laboral e ingresos. Muestran la estabilización de los valores a medida que continúa el reclutamiento, lo que sugiere que la muestra resultante no está sesgada por la selección inicial de semillas. Se han obtenido gráficos similares y resultados similares para todas las demás características en estudio y para los estimadores RDS I y RDS II, pero no se informan aquí para facilitar la presentación.

La Figura 3.4 muestra los gráficos de cuello de botella, que parecen converger en una estimación puntual para cada variable, lo que sugiere estimaciones estables (en lugar de converger en dos o tres, lo que indicaría estimaciones inestables y diferencias importantes entre los datos de diferentes semillas). En Gile, Johnston y Salganik [35] se encuentran disponibles ejemplos de gráficos de cuello de botella. Calculamos estimaciones de RDS II y RDS SS para cada variable de la encuesta. Las diferencias entre los estimadores RDS II y RDS SS son muy pequeñas (por debajo de 0,01) para todas las variables de la encuesta (consulte la tabla 1 del Apéndice .1), lo que indica que el tamaño de la población no induce sesgo en nuestras estimaciones [35].

Tabla 3.4: Estimaciones RDS e intervalos de confianza del 95 % para jóvenes urbanos indígenas, monubios y afroecuatorianos en el cantón de Riobamba (n = 814) y datos oficiales para esta etnia y para los ecuatorianos habituales en Riobamba según ENEMDU 2018.

Característica	**Categoría**	**Total ENEMDU**	**ENEMDU Minoría Étnica**	**Estimación RDS**	**Intervalo Confianzaal 95 %**		**Efecto Diseño**	**SD**	**n**
Sexo	Hombre	0.4575	0.497	0.4261	0.347	0.5051	5.8066	0.0403	364
	Mujer	0.5425	0.503	0.5739	0.4949	0.653	5.8066	0.0403	450
	Edad	23.1970 (3.6322)	23.2482 (3.4481)	21.8133	21.3022	22.3243	5.7469	0.2607	814
Estado civil	Casado-Unión libre	0.2310	0.3896	0.1827	0.133	0.2323	3.7594	0.0253	163
	Divorciado-separado	0.0310	0.0576	0.037	0.0015	0.0058	0.2806	0.0011	4
	Soltero	0.7380	0.5509	0.8131	0.7632	0.863	3.7236	0.0254	646
	Viudo	0	0.0019	0.0006	0.0003	0.0008	0.0316	0.0001	1
Autoidentificación étnica	Afroecuatoriano	0.0586 *		0.0504	0.0324	0.0683	1.532	0.0092	44
	Indígena	0.9256 *		0.9197	0.8804	0.959	4.762	0.0201	749
	Montubia	0.0158 *		0.0299	0	0.0658	10.0818	0.0183	21
Instrucción	Secundaria o menos	0.5450	0.8179	0.7985	0.7546	0.8424	2.7214	0.0224	624
	Superior	0.4550	0.1821	0.2015	0.1576	0.2454	2.7214	0.0224	190
Lenguaje	Español	0.8422	0.7672	0.7585	0.6918	0.8253	5.5309	0.0341	608
	Indígena	0.1578	0.2328	0.2415	0.1747	0.3082	5.5309	0.0341	206
Trabajo	No	0.5372	0.4398	0.5009	0.4216	0.5801	5.7114	0.0404	375
	Si	0.4628	0.5602	0.4991	0.4199	0.5784	5.7114	0.0404	439
Seguridad Social	Asegurado	0.3160	0.1535	0.1505	0.0983	0.2026	2.5221	0.0266	71
	No asegurado	0.6840	0.8465	0.8495	0.7974	0.9017	2.5221	0.0266	368
Salario	Salario mensual	523.58 (381.61)	379.64 (267.43)	295.496	259.599	331.393	1.7204	18.3148	439
Pobreza extrema	Si	0.0803	0.0629	0.0935	0.0506	0.1364	4.8907	0.0219	71
	No	0.9197	0.9371	0.9065	0.8636	0.9494	4.8907	0.0219	737
Servicio de energia	Compañía publica	0.9861		0.9524	0.9345	0.9704	1.6166	0.0092	774
	Generador, vela u otro	0.0139		0.0476	0.0296	0.0655	1.6166	0.0092	40
Servicio de agua	Pozo-grieta, granel u otro	0.0113		0.0793	0.0539	0.1047	2.0048	0.0129	66
	Tubería	0.9887		0.9207	0.8953	0.9461	2.0048	0.0129	748
Hogar	Número de cuartos	2.69 (1.06)		2.8083	2.6214	2.9952	5.9538	0.0954	814
	Número de personas	2.96 * (2.54 *)		4.6901	4.4354	4.9448	3.7184	0.13	814
Percepción de discriminación	Víctima	-		0.9063	0.8633	0.9493	49493	0.0219	727
	Satisfacción general	-		8.5506	8.2982	8.8029	6.7684	0.1288	814

* Estimaciones utilizando el CPV 2010.
Los números en color azul son de aquellas estimaciones RDS SS, para las cuales el valor oficial para ecuatorianos regulares en Riobamba (Total ENEMDU) está fuera del intervalo RDS SS al 95 % de confianza.
Los números en color verde son de aquellas estimaciones de RDS SS, para las cuales el valor oficial para ecuatorianos pertenecientes a minorías étnicas en Riobamba (ENEMDU Minoría Étnica) está fuera del intervalo RDS SS al 95 % de confianza.

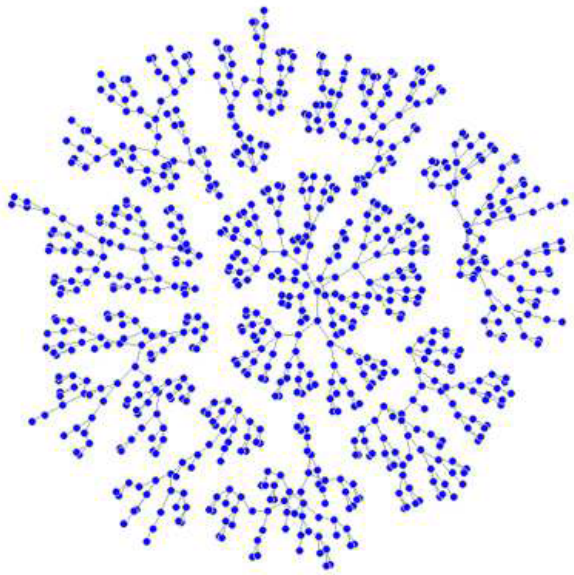

Figura 3.1: Representación de la red de cadenas de reclutamiento para RDS de jóvenes urbanos de minorías étnicas en el cantón de Riobamba, Ecuador.

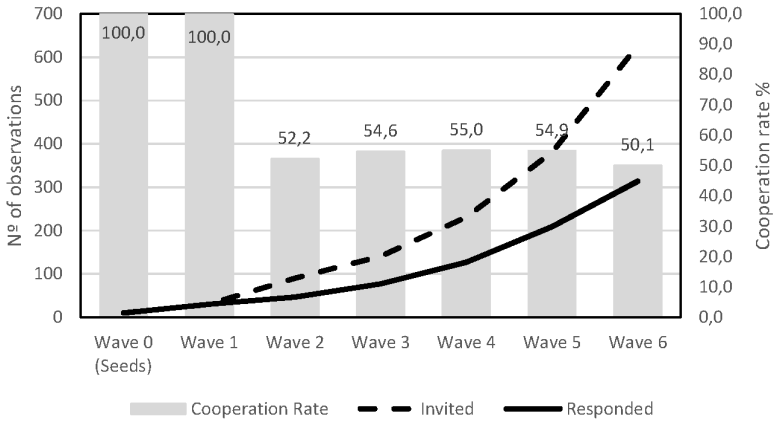

Figura 3.2: Número de invitaciones, número de participantes y tasas de cooperación por oleada de trabajo de campo RDS.

Figura 3.3: Gráficos de convergencia que muestran las observaciones de la muestra con el rasgo seleccionado: (**a**) mujeres, (**b**) secundaria o menos, (**c**) salario y (**d**) trabajo. La línea discontinua muestra la estimación basada en la muestra completa.

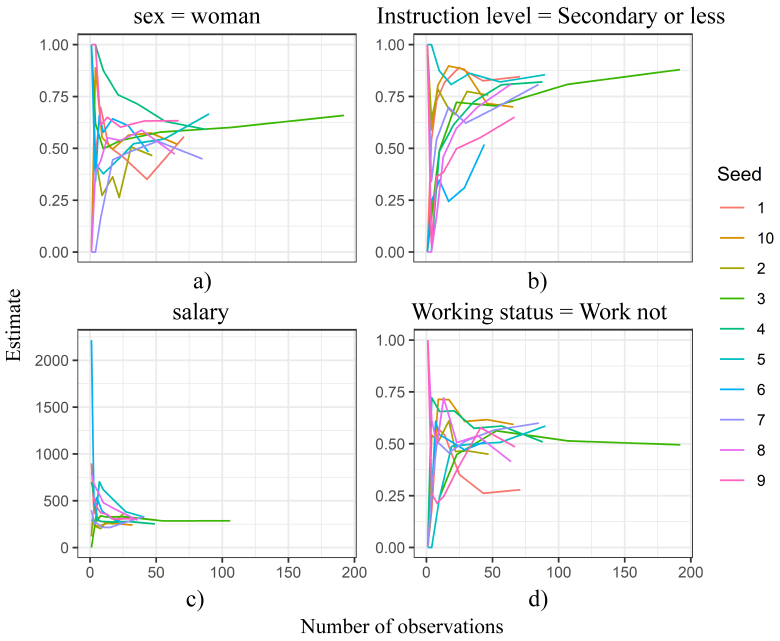

Figura 3.4: Gráficos de cuello de botella que muestran las observaciones en cada cadena de las semillas con el rasgo seleccionado: (**a**) mujer, (**b**) secundaria o menos, (**c**) salario y (**d**) trabajo.

Capítulo 4

Regresión en RDS

4.1. Introducción

La investigación actual sobre RDS se centra principalmente en estimar las medias y proporciones de la población [23, 35]. El rendimiento de los estimadores de prevalencia de RDS se ha evaluado en muchos estudios, utilizando diferentes métodos, en particular simulaciones (por ejemplo, [43, 104]), con resultados diferentes. En general, los estudios muestran un rendimiento intermedio a alto de los estimadores de prevalencia de RDS [104, 105]. Por lo tanto, casi todos los estimadores propuestos actualmente para muestras de RDS tienen como objetivo estimar la prevalencia de una condición en la población de interés y no la identificación de factores asociados con esa condición.

En el mismo sentido, Gile et al. [23] destacan que los métodos para el modelado multivariable están poco desarrollados a pesar de tener una gran demanda. Esto se refleja en la variedad de enfoques analíticos adoptados en la literatura aplicada. Algunos estudios tratan los datos de RDS como si fueran recolectados por muestreo aleatorio y aplican ANOVA, regresiones lineales y logísticas sin ningún ajuste para los pesos de muestreo de RDS [24]. Otros incluyen ponderaciones RDS en modelos de regresión, basándose en la suposición típica de RDS que es más probable que algunos individuos sean reclutados en la muestra que otros [25]. Los métodos que incorporan ponderaciones muestrales tienden a mejorar su desempeño cuando la homofilia es pequeña en la población, ya que por lo general no tienen en cuenta la dependencia potencial de las unidades. Si bien esto puede ser un problema en poblaciones con alta homofilia, no existe

un método de regresión claro y confiable en RDS que contabilice el agrupamiento y que pueda usarse ampliamente en aplicaciones. El ajuste por clústeres requiere conocimiento previo de la población y si no se realiza bien en la práctica, o incluso si los clústeres no existen realmente en la población, puede resultar en estimaciones sesgadas [106]. Hubbart et al. [107] señalan algunos de los problemas relacionados con el uso de ecuaciones de estimación generalizadas (GEE, por sus siglas en inglés: generalized estimating equations) cuando el número de clústeres es pequeño y Rao et al. [106] enfatizan algunos de los problemas relacionados con un ajuste incorrecto de los clústeres. En otro enfoque, Rhodes y McCoy [26] incluyen semillas como efectos aleatorios para ajustar la dependencia dentro de las cadenas de reclutamiento, pero ignoran los pesos de RDS. Una consulta de la base de datos Pubmed identifica 70 estudios publicados entre 2018 y 2019 que aplican modelos de regresión logística a muestras de RDS, con un 48,6 % (n = 34) utilizando estimadores no ponderados y un 44,3 % (n = 31) utilizando alguna forma de ponderación con grados de red y el 7,1 % (n = 5), que presentan modelos ponderados y no ponderados. Este patrón pone de relieve la evidente falta de consenso en la literatura actual sobre qué tipo de estimador debe utilizarse [108].

Por lo tanto, si bien existen estrategias bien desarrolladas para estimar medias y prevalencias a partir de estudios de RDS, el uso de estimadores basados en modelos para estudiar las relaciones entre la respuesta y las variables explicativas no ha sido suficientemente investigado, especialmente en lo que respecta a la pregunta básica de cuándo usar ponderaciones muestrales [30].

En la sección 4.2 se exponen las preocupaciones en la aplicación de regresión con datos RDS y se describe el análisis previo a la construcción del modelo de regresión. La sección 4.3 describe dos estrategias para llevar a cabo un análisis de regresión con datos RDS. La sección 4.4 muestra una de las aportaciones fundamentales del presente trabajo, que es la estimación de parámetros no lineales con datos RDS, considerando el problema del modelado de regresión y la asociación de datos RDS.

4.2. Preliminares en el modelado RDS

4.2.1. Preocupaciones en el modelado RDS

En esta sección, siguiendo a Spiller [27] se describen las principales preocupaciones en la aplicación de regresión en datos recolectados mediante la metodología RDS, es decir, se describe la diferencia entre

datos RDS y los obtenidos mediante muestreo tradicional.

La principal preocupación de los modeladores de datos RDS es adaptarse a la falta de independencia entre los encuestados. Los modelos de regresión estándar asumen que los errores a nivel individual no están correlacionados con las variables independientes del modelo (lo que implica que las observaciones se muestrean independientemente de la población). Dado que algunos encuestados de la muestra de RDS reclutan a más de un encuestado, esta suposición no es válida para los datos de RDS. En este caso, se tratan a los encuestados que comparten un reclutador como un clúster.

El proceso de reclutamiento sugiere que la dependencia entre las observaciones es más fuerte en el nivel de díada (reclutador-reclutado), pero la mayoría de los encuestados son reclutados y reclutadores, por lo que las díadas no forman clústeres mutuamente excluyentes (lo que se requiere para la mayoría de las estrategias de ajuste de regresión). En la mayoría de los estudios de RDS, los encuestados pueden realizar más de un reclutamiento, por lo que deberíamos esperar observar un agrupamiento fuerte por reclutador compartido si las redes sociales de los reclutadores son más homogéneas que la red social completa que se muestrea. Dado que la red de población consta de clústeres sociales reales (es decir, iglesias, vecindarios, clubes, etc.), una preocupación central de la estrategia de ajuste es la agrupación no observable en la red social de la que RDS está tomando muestras.

La segunda gran preocupación para los modeladores de datos RDS es que la información necesaria para obtener estimaciones de la población a partir de datos de muestra (pero no para obtener estimaciones de varianza, donde se necesita la probabilidad de inclusión de segundo orden) es capturada completamente por la probabilidad de inclusión de los encuestados en la muestra. En RDS, esta probabilidad de inclusión está influida por dos componentes: el número de reclutadores potenciales que un encuestado conoce en la población objetivo y las características de reclutamiento del clúster en una variable específica (por ejemplo, las características de los afroamericanos en una muestra para la variable etnia). El primer componente es estable en todas las variables, pero el segundo no lo es. Afortunadamente, existe una gran cantidad de bibliografía a partir de la cual se puede aprovechar el modelado de datos agrupados (o correlacionados) y la estimación de encuestas que se ajustan a las probabilidades de inclusión. Desafortunadamente, esta literatura está orientada casi universalmente hacia el muestreo clásico de encuestas de múltiples etapas.

4.2.2. Exploración de datos RDS

El analista, antes de la construcción del modelo debe identificar la estructura de la red subyacente. Para esto en primer lugar, se evalúa mediante el cálculo de la homofilia utilizando el enfoque de estimación estándar de RDS y teniendo en cuenta si la muestra se ha mezclado en el área geográfica y los sitios de entrevistas. Si la red subyacente está integrada geográficamente, es de esperar ver el área geográfica (y el sitio de la entrevista) distribuida aleatoriamente dentro y entre los árboles de reclutamiento. Si la homofilia del área geográfica es alta, el modelador debe considerar incluir la geografía como un factor en su modelo de regresión. Si hay más de un sitio en cualquier área geográfica, el modelador debe examinar la homofilia por sitio para asegurarse que la muestra no esté segregada por sitio dentro del área (es decir, para asegurarse que la red está verdaderamente estructurada por área geográfica y no en un nivel más fino). Si no hay mezcla entre sitios, la muestra debe dividirse en múltiples subconjuntos para la estimación de la población (como en Heckathorn [4]), lo que evita los problemas planteados por la suposición que la red de la población forma un componente conectado, es decir, que existe un camino entre cada persona y todas las demás.

En segundo lugar, es importante examinar las variables o características según las cuales los encuestados se clasifican, mediante el cálculo de la homofilia. En poblaciones bien mezcladas y homogéneas, es posible observar muy poca o ninguna homofilia, lo que simplifica el trabajo del analista. Debido a que la homofilia solo es relevante para la estimación consistente de coeficientes en el modelo si está relacionada con la variable de resultado, la homofilia general para cada variable es menos informativa que la homofilia por la variable de resultado. Por ejemplo, si el resultado del modelo es si un encuestado es VIH positivo o no, el modelador debe examinar la homofilia racial según el estado del VIH en lugar de la homofilia racial en general.

En tercer lugar, el analista debe evaluar el grado en que la muestra revela los clústeres sociales que forman la red que RDS está muestreando. Observar directamente la pertenencia a estos clústeres es virtualmente imposible, porque un investigador no podría construir una lista de clústeres a priori.

Debido a la naturaleza estocástica del proceso de muestreo (principalmente debido al supuesto de reclutamiento aleatorio), el grado en que la muestra se mapea en los clústeres sociales subyacentes también es un proceso estocástico (es decir, si se muestrea la misma población repetidamente, el grado en que la muestra revela clústeres sociales sería un resultado estocástico basado en una distribución muestral po-

blacional desconocida). A la luz de esto, el analista debe examinar el grado de similitud de los encuestados (correlación intraclase) en diferentes niveles de agrupación observable. El nivel más bajo de agrupación se encuentra entre los reclutados que comparten un reclutador; el nivel más alto de agrupamiento es por árbol de reclutamiento. El analista debe examinar el agrupamiento en ambos niveles y anticipar el ajuste del modelo de regresión si se detecta un agrupamiento significativo.

4.3. Modelado de regresión con datos RDS

De nuestro conocimiento, existe poca investigación sobre la estimación de factores de riesgo para poblaciones de difícil acceso teniendo en cuenta el enfoque RDS, es decir, considerando la homofilia inherente al proceso de reclutamiento y a la agrupación de redes subyacentes a la muestra. A continuación, describimos dos estrategias para realizar análisis de regresión para datos RDS desarrolladas recientemente.

4.3.1. Regresión binaria

Bastos et al. [109] proponen un modelo de regresión binaria teniendo en cuenta la estructura RDS, que se incluye en el modelo mediante un efecto aleatorio latente con una estructura de correlación. Este modelo se describe a continuación:

Sea y_i una variable que representa una característica de interés del i-ésimo individuo entrevistado en una muestra RDS, donde $y_i = 1$ si la característica de interés se observa en el individuo i y $y_i = 0$ en caso contrario para $i = 1, 2, \ldots, n$. Los factores de riesgo se pueden incorporar en un modelo de regresión binaria de la siguiente manera

$$y_i \sim Bernoulli(\vartheta_i),\ g(\vartheta_i) = \eta_i = \mathbf{x}_i^T \beta,\ i = 1, 2, \ldots, n \tag{4.1}$$

donde $\mathbf{x}_i$ es un vector de posibles factores de riesgo, β son los efectos de riesgo y $g(.)$ es una función de enlace. Si la función de enlace es la función logit, $g(z) = logit(z) = log(z/(1-z))$, entonces la regresión se llama regresión logística.

Sin embargo, el modelo (4.1) es válido solo cuando la característica de interés es independiente entre los individuos en un estudio de RDS, es decir, esto es válido cuando no hay dependencia de la red.

Si se conoce la red de contactos, entonces se puede incluir un término latente en el modelo logístico

donde se tendrá en cuenta la estructura de la red. Esto se hace usando un modelo aleatorio Gaussiano de Markov latente, es decir,

$$y_i \sim Bernoulli(\vartheta_i),\ g(\vartheta_i) = \eta_i = \mathbf{x}_i^T\beta + \varpi_i,\ i = 1, 2, \ldots, n \tag{4.2}$$

donde ϖ_i es un efecto latente de la estructura de la red. Los efectos latentes se modelan utilizando el siguiente modelo autorregresivo condicional (CAR, por sus siglas en inglés: conditional auto-regressive), propuesto por Besag [110],

$$\varpi_i|\varpi_j, i \neq j, d, \iota \sim N\left(\frac{1}{d+n_i}\sum_{i\sim j}\varpi_j, \frac{1}{(d+n_i)\iota}\right), \tag{4.3}$$

n_i es el número de contactos del individuo i (número de conexiones), $i \sim j$ significa el conjunto de individuos conectados a i, ι es un hiperparámetro de precisión y d es un parámetro diagonal. Para completar el modelo, se establecen a priori valores para β, ι y d.

El modelo (4.3) es un modelo bien conocido en Estadística Espacial Bayesiana, donde las regiones vecinas se consideran conexiones [111]. La inferencia se basa en las distribuciones marginales posteriores de cada parámetro. Estas distribuciones marginales posteriores se obtienen utilizando la aproximación de Laplace anidada integrada [112].

Finalmente, Bastos et al. [109] indican que el modelo de regresión binaria con efectos latentes puede ser una alternativa a los modelos de regresión para datos RDS que ignoran la estructura de la red. Además, se establece que la regresión binaria con el efecto de red latente asume que la red observada contiene toda la información sobre la red social. Sin embargo, la red observada a partir de los datos de muestreo de los encuestados está incompleta. Esto se debe al número limitado de contactos que puede traer cada persona y también al hecho de que cada individuo no puede participar más de una vez en el estudio.

4.3.2. Regresión general

Más recientemente, Yauck et al. [28] proponen una metodología para técnicas de regresión general utilizando datos RDS. En el trabajo describen que una muestra de RDS tiene una estructura gráfica, que típicamente es una red social parcialmente observada de individuos reclutados con una estructura de dependencia subyacente desconocida, en la que es común observar una tendencia de individuos con

rasgos similares a compartir vínculos sociales, una característica denominada homofilia. Además, los autores indican que el proceso de RDS no es puramente aleatorio, sino que es más probable que algunos individuos sean seleccionados en la muestra que otros. Un principio subyacente asumido en RDS es que la probabilidad que un individuo sea reclutado depende del tamaño de su red personal de contactos sociales [4, 44].

En la metodología propuesta por Yauck et al. [28] se modela conjuntamente los efectos debidos a la homofilia y la dependencia entre los resultados de los clústeres de la red de población no observada. Esto permite ver el ajuste del modelo supuesto para los datos RDS observados, como un problema de datos faltantes.

Estructura de una red RDS

En primer lugar, se define la estructura de la red resultante de un muestreo RDS. Este muestreo se desarrolla en una población infinita en la que los individuos están conectados por lazos sociales, considerando tres supuestos

1. Red de la población: La red de la población representa un número infinito de clústeres de tamaños finitos que no se superponen. Esto es, la población está agrupada, con los individuos distribuidos en clústeres.

2. Reclutamiento de RDS. El proceso de reclutamiento se lleva a cabo dentro de un subconjunto de clústeres de la red y progresa a través de las conexiones sociales de las personas.

3. Sin reclutamientos múltiples. Ningún individuo puede ser reclutado más de una vez para el estudio.

Los tres supuestos anteriores implican que la red RDS observada se puede representar como un conjunto de árboles no superpuestos.

Modelo de regresión

Sea y_{ij} el resultado del j-ésimo individuo del i-ésimo clúster, $j = 1, \ldots, N_i$, donde N_i es el tamaño del i-ésimo clúster, $i = 1, \ldots, m$. Sea x_{ij} el valor de la covariable para el j-ésimo individuo del i-ésimo clúster y $\mathbf{x}_i$ el vector de covariables para todos los individuos del i-ésimo clúster. Suponemos que $\{y_{ij}, x_{ij}; i = 1, \ldots, m; j = 1, \ldots, N_i\}$ es la realización de una muestra aleatoria cuya distribución es

idéntica a la superpoblación de clústeres definida en el apartado 4.3.2 (Estructura de una red RDS), por lo que cualquier inferencia basada en la muestra corresponde a los parámetros de la población infinita de la que se extrae la muestra. Los autores, siguiendo a Manski [113] asumieron que la relación subyacente entre el resultado y las covariables en la población se caracteriza por un modelo lineal mixto generalizado en el que $\boldsymbol{\delta}_i = (\delta_{i1} \ldots, \delta_{iN_i})$ es un vector de efectos aleatorios para el i-ésimo clúster, $\mu_{ij} = E(y_{ij}|\mathbf{x}_i, \delta_{ij})$ y

$$g(\mu_{ij}) = \beta_0 + \beta_1 x_{ij} + \gamma \frac{1}{n_{ij}} \sum_{k \sim j} x_{ik} + \delta_{ij}, \tag{4.4}$$

donde $g(.)$ es una función (monótona) de la media, $k \sim j$ representa el conjunto de individuos que comparten lazos con el j-ésimo individuo, n_{ij} es el número de conexiones sociales que el j-ésimo individuo del i-ésimo clúster comparte con otros individuos dentro del mismo clúster o grado. Suponemos además que $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \sum_i)$, con $cov(\boldsymbol{\delta}_i, \boldsymbol{\delta}_j) = \mathbf{0}$ para $i \neq j$. El parámetro γ mide los efectos debidos a la homofilia, o la influencia de las características de los contactos en el resultado de un individuo. En este modelo, los parámetros β_0 y (el parámetro potencialmente con valor vectorial) β_1 son de interés principal.

Luego se define $\boldsymbol{S}^{(i)} = (s_{jk}^{(i)})$ como una matriz (de vecindad) que representa los lazos sociales en el i-ésimo clúster tal que $s_{jk}^{(i)} = 1$ si el individuo j y el individuo k comparten un vínculo y $s_{jk}^{(i)} = 0$ de lo contrario, con $s_{jj}^{(i)} = 0$ y $\boldsymbol{S} = diag(\boldsymbol{S}^{(i)})$. Se supone un modelo autorregresivo simultáneo (SAR, por sus siglas en inglés: simultaneous auto-regressive model) [114, 115] para el vector de efectos aleatorios $\boldsymbol{\delta}_i$:

$$\boldsymbol{\delta}_i = \rho \boldsymbol{S}^{(i)} \boldsymbol{\delta}_{i-1} + \boldsymbol{\mu}_i, \tag{4.5}$$

donde ρ representa la fuerza de la dependencia dentro de la red y $\boldsymbol{\mu}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{N_i})$. Dado que $\boldsymbol{W}_i = (\boldsymbol{I}_{N_i} - \rho \boldsymbol{S}^{(i)})^{-1}$ existe, la covarianza de $\boldsymbol{\delta}_i$, $\sum_i$ se puede escribir como

$$\sum\nolimits_i (\sigma^2, \rho) = \sigma^2 \boldsymbol{W}_i \boldsymbol{W}_i^T. \tag{4.6}$$

La matriz de correlación SAR es tal que los resultados de los individuos vecinos (es decir, socialmente conectados) están más correlacionados que los resultados de los no vecinos. Otros modelos de correlación para $\boldsymbol{\delta}_i$ con tales propiedades, incluyen los modelos autorregresivos condicionales CAR, que pertenecen a la misma clase de modelos de área que los modelos SAR y modelos que asumen una función de correlación

que depende de la distancia entre observaciones [116].

Yauck et al. [28] concluyen que el desarrollo de métodos de regresión para RDS está limitado por un problema de datos faltantes, ya que los datos RDS observados revelan solo información parcial sobre la red RDS completa y no observada. En este caso, no se puede realizar una inferencia válida sobre los efectos impulsados, por la homofilia y/o las estructuras de correlación inducidas por la red sin datos de red adicionales o restricciones topológicas estrictas en la red RDS [117]. También, mostraron que ignorar los efectos impulsados por la homofilia, si están presentes, induce un sesgo de pequeño a insignificante en el estimador de parámetros (de la covariable homofílica) para los modelos lineales y de Poisson, mientras que induce un sesgo sustancial para la regresión logística cuando se asume la agrupación en el nivel de semilla. Además, especificar incorrectamente el modelo de correlación induce un sesgo creciente a medida que aumenta la dependencia dentro de la red RDS y una cobertura deficiente para los intervalos de confianza basados en el modelo. Igualmente, se muestra que los métodos de regresión ponderados superan a los métodos de regresión no ponderados en términos de sesgo cuando el predictor está correlacionado con el grado, asumiendo que no falta ninguna covariable en el modelo. La ponderación del modelo solo agrega variabilidad en las estimaciones cuando el predictor y el resultado no están correlacionados.

4.4. Estimación de parámetros no lineales con datos RDS

En este apartado, se presenta una de las aportaciones más importantes de esta tesis: la estimación de parámetros no lineales con datos RDS, donde se considera el problema del modelado de regresión y asociación para datos RDS continuos. Publicado recientemente en la revista Mathematics (Sánchez-Borrego, I.; Rueda, M.; Mullo, H. Estimation of Non-Linear Parameters with Data Collected Using Respondent-Driven Sampling. *Mathematics.* **2020**, *8(8)*, 1315, doi:10.3390/math8081315.). Factor de impacto (2019): 1.747. Posición 28/325 (Q1) en el listado JCR Mathematics - SCIE.

4.4.1. Introducción

El método propuesto en esta sección aborda el problema del modelado de regresión y la asociación entre variables continuas al proponer un nuevo método de estimación de los pesos muestrales para datos continuos. El enfoque de nuestro trabajo es proponer un método para estimar parámetros no lineales como la covarianza y el coeficiente de correlación. Se proponen estimadores que hacen uso de estimadores

habituales en RDS que contemplan datos continuos (RDS II y RDS SS) y se demuestra que, al igual que ellos, presentan propiedades como la insesgadez asintótica. En la Figura 4.1 se muestra un diagrama que ilustra este método. Se proponen además estimadores de las varianzas de los métodos propuestos. Nuestro método se centra en la asociación bivariada entre variables continuas y pretende cubrir la necesidad de estudios como estos en el contexto de RDS, ya que la mayoría de los estudios incorporan las ponderaciones utilizando software estadístico estándar y, a diferencia de nuestra propuesta, se centran en la estimación de la prevalencia.

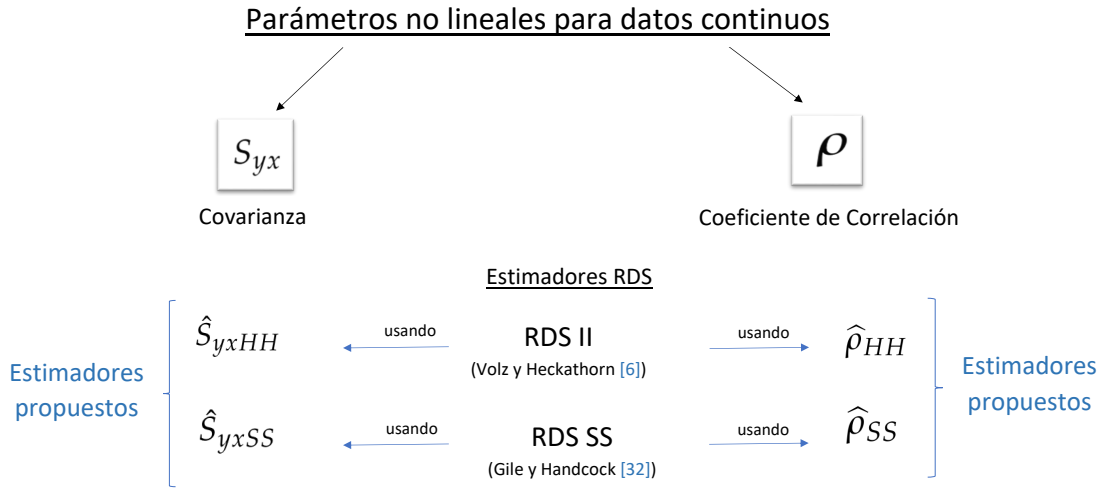

Figura 4.1: Representación esquemática del método propuesto.

En la siguiente sección 4.4.2 se proponen estimadores para la covarianza poblacional y para el coeficiente de correlación. Los estimadores de las varianzas de los estimadores propuestos se consideran en la sección 4.4.4. Se realiza un estudio de simulación para estudiar su comportamiento en la práctica en la sección 4.4.5. En la sección 4.4.6 se presenta una ilustración del método para estudiar las condiciones de vida de los jóvenes indígenas, monubios y afroecuatorianos.

4.4.2. Estimación de algunos parámetros no lineales

El uso generalizado de la regresión basada en datos provenientes de encuestas por muestreo requiere una evaluación previa y cuidadosa del uso de técnicas clásicas. Es evidente que los estimadores habituales de los parámetros implicados en la regresión no son válidos en el contexto RDS. En esta sección, desa-

rrollamos algunos estimadores para varianzas poblacionales, covarianzas y el coeficiente de correlación.

Estimación de la varianza y covarianza

Consideramos la covarianza poblacional como:

$$S_{yx} = \frac{1}{N-1}\sum_U (y_k - \overline{y})(x_k - \overline{x}),$$

donde y es la variable de interés y x es la covariable (o variable auxiliar), ambas variables de tamaño N, $\overline{y}$ y $\overline{x}$ es la media de y y x correspondientemente.

Podemos escribir este parámetro como:

$$S_{yx} = \frac{1}{N-1}T_{yx} - \frac{1}{N(N-1)}T_y T_x = \theta = f(\theta_1, \theta_2, \theta_3),$$

siendo $T_{yx} = \theta_1 = \sum_U y_k x_k$, $T_y = \theta_2 = \sum_U y_k$ y $T_x = \theta_3 = \sum_U x_k$.

De manera similar, las varianzas de población finita se definen como

$$S_y^2 = \frac{1}{N-1}T_{yy} - \frac{1}{N(N-1)}T_y^2,$$

y

$$S_x^2 = \frac{1}{N-1}T_{xx} - \frac{1}{N(N-1)}T_x^2.$$

Se construyen estimadores para estos parámetros asumiendo que y_k y x_k se observan para las unidades de la muestra s de RDS.

Si existen $\hat{\theta}_1$, $\hat{\theta}_2$ y $\hat{\theta}_3$ estimadores consistentes de θ_1, θ_2 y θ_3, un estimador consistente de S_{yx} será

$$\hat{S}_{yx} = \frac{1}{N-1}\hat{\theta}_1 - \frac{1}{N(N-1)}\hat{\theta}_2\hat{\theta}_3 = \hat{\theta}. \tag{4.7}$$

Podemos estimar estos totales con el estimador RDS II:

$\hat{T}_{yHH} = \frac{N\hat{\delta}_v}{n}\sum_s y_k\delta_k^{-1}$, $\hat{T}_{yyHH} = \frac{N\hat{\delta}_v}{n}\sum_s y_k^2\delta_k^{-1}$, y $\hat{T}_{yxHH} = \frac{N\hat{\delta}_v}{n}\sum_s y_k x_k\delta_k^{-1}$, siendo $\hat{\delta}_v = \frac{n}{\sum_U \delta_k^{-1}}$ el grado medio.

Entonces, el estimador de la covarianza es

$$\hat{S}_{yxHH} = \frac{1}{N-1}\hat{T}_{yxHH} - \frac{1}{N(N-1)}\hat{T}_{yHH}\hat{T}_{xHH}. \tag{4.8}$$

Si N es grande, $\hat{T}_{yHH}$ se puede escribir de una manera más sencilla que no depende de N:

$$\hat{S}_{yxHH} = \frac{\hat{\delta}_v}{n}\sum_s y_k x_k \delta_k^{-1} - \frac{\hat{\delta}_v^2}{n^2}\sum_s y_k \delta_k^{-1}\sum_s x_k \delta_k^{-1}. \tag{4.9}$$

Utilizando la idea del estimador RDS SS, proponemos estimar los totales como:

$\hat{T}_{ySS} = \sum_s y_k \hat{\pi}(\delta_k)^{-1}$, $\hat{T}_{yxSS} = \sum_s y_k x_k \hat{\pi}(\delta_k)^{-1}$ y $\hat{T}_{yySS} = \sum_s y_k^2 \hat{\pi}(\delta_k)^{-1}$, siendo $\hat{\pi}(\delta_k)$ la distribución poblacional estimada de grados a través de muestreos sucesivos.

Entonces, el estimador de la covarianza es

$$\hat{S}_{yxSS} = \frac{1}{N-1}\hat{T}_{yxSS} - \frac{1}{N(N-1)}\hat{T}_{ySS}\hat{T}_{xSS}.$$

Si se desconoce N, un estimador consistente para S_{yx} es:

$$\hat{S}_{yxSS} = \frac{1}{\hat{N}-1}\hat{T}_{yxSS} - \frac{1}{\hat{N}(\hat{N}-1)}\hat{T}_{ySS}\hat{T}_{xSS}, \tag{4.10}$$

con $\hat{N} = \sum_s \hat{\pi}(\delta_j)^{-1}$.

Los estimadores RDS SS y RDS II de un total son asintóticamente insesgados, por lo que los estimadores propuestos son asintóticamente insesgados.

4.4.3. Estimación del coeficiente de correlación

En esta sección, consideramos la estimación del coeficiente de correlación entre dos variables, y y x, definido por

$$\rho = S_{yx}/S_y S_x.$$

Se pueden obtener dos estimadores para este parámetro utilizando los estimadores RDS II y RDS SS previamente definidos:

$$\widehat{\rho}_{HH} = \frac{\hat{S}_{yxHH}}{\hat{S}_{yHH}\hat{S}_{xHH}}, \tag{4.11}$$

y

$$\widehat{\rho}_{SS} = \frac{\hat{S}_{yxSS}}{\hat{S}_{ySS}\hat{S}_{xSS}}, \tag{4.12}$$

siendo $\hat{S}_{yHH} = \frac{1}{N-1}\hat{T}_{yyHH} - \frac{1}{N(N-1)}\hat{T}^2_{yHH}$, $\hat{S}_{xHH} = \frac{1}{N-1}\hat{T}_{xxHH} - \frac{1}{N(N-1)}\hat{T}^2_{xHH}$, $\hat{S}_{ySS} = \frac{1}{N-1}\hat{T}_{yySS} - \frac{1}{N(N-1)}\hat{T}^2_{ySS}$ y $\hat{S}_{xSS} = \frac{1}{N-1}\hat{T}_{xxSS} - \frac{1}{N(N-1)}\hat{T}^2_{xSS}$.

4.4.4. Estimación de las varianzas

Consideramos la estimación de la varianza de la covarianza de $\widehat{S}_{yx}$.

Usando una linealización de Taylor, escribimos

$$\widehat{\theta} \simeq \widehat{\theta}_o = \theta + \sum_1^3 w_j(\widehat{\theta}_j - \theta_j),$$

con

$$w_j = \frac{\partial f(\widehat{\theta}_1(s), ..., \widehat{\theta}_3(s))}{\partial \widehat{\theta}_j}|_{\theta_1,...,\theta_3}.$$

$$V(\widehat{\theta}_o) = V(\sum w_j\widehat{\theta}_j) = \sum w_j^2 V(\widehat{\theta}_j) + \sum w_i w_j cov(\widehat{\theta}_i, \widehat{\theta}_j),$$

y

$$\widehat{V}(\widehat{\theta}) \simeq \widehat{V}(\widehat{\theta}_o) = \sum \hat{w}_j^2 \widehat{V}(\widehat{\theta}_j) + \sum \hat{w}_i\hat{w}_j\widehat{cov}(\widehat{\theta}_i, \widehat{\theta}_j),$$

siendo $w_1 = \frac{1}{N-1}$, $w_2 = -\frac{\theta_3}{N(N-1)}$, $w_3 = -\frac{\theta_2}{N(N-1)}$.

Estos pesos son estimados por

$$\hat{w}_1 = w_1,\ \hat{w}_2 = -\frac{\hat{T}_x}{N(N-1)},\ \hat{w}_3 = -\frac{\hat{T}_y}{N(N-1)}.$$

Una expresión computacional más sencilla se puede obtener a partir de las fórmulas 5.5.10 en Särndal et al. [118].

Se estiman las varianzas y covarianzas de los totales antes mencionados para el estimador RDS II como

$$\hat{V}(\hat{T}_{yxHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v y_k x_k \delta_k^{-1} - \hat{T}_{yxHH})^2,$$

$$\hat{V}(\hat{T}_{yHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v y_k \delta_k^{-1} - \hat{T}_{yHH})^2,$$

$$\hat{V}(\hat{T}_{xHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v x_k \delta_k^{-1} - \hat{T}_{xHH})^2,$$

$$\widehat{cov}(\hat{T}_{yxHH}, \hat{T}_{xHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v y_k x_k \delta_k^{-1} - \hat{T}_{yxHH})(N\hat{\delta}_v x_k \delta_k^{-1} - \hat{T}_{xHH}),$$

$$\widehat{cov}(\hat{T}_{yxHH}, \hat{T}_{yHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v y_k x_k \delta_k^{-1} - \hat{T}_{yxHH})(N\hat{\delta}_v y_k \delta_k^{-1} - \hat{T}_{yHH}),$$

y

$$\widehat{cov}(\hat{T}_{yHH}, \hat{T}_{xHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v y_k \delta_k^{-1} - \hat{T}_{yHH})(N\hat{\delta}_v x_k \delta_k^{-1} - \hat{T}_{xHH}).$$

El estimador RDS II propuesto es sólo análogo al estimador de Hansen y Hurvitz [119], pero como los datos están correlacionados en un marco RDS, los estimadores mencionados anteriormente pueden funcionar mal. A pesar que Volz y Heckathorn [6] obtuvieron un estimador de varianza que considera la estructura MCMC de la muestra para variables categóricas, no podemos usar este estimador de varianza en este contexto.

Se estima ahora las varianzas y covarianzas de los totales para el estimador RDS SS utilizando el método de Deville [120] para estimar la varianza del estimador de Horvitz-Thompson. Las variaciones se estiman como

$$\widehat{V}(\widehat{T}_{yxSS}) = \frac{1}{1-\sum_{k\in s} a_k^2}\sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{y_k x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l x_l/\hat{\pi}(\delta_l))^2,$$

$$\widehat{V}(\widehat{T}_{ySS}) = \frac{1}{1-\sum_{k\in s} a_k^2}\sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{y_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l/\hat{\pi}(\delta_l))^2,$$

y

$$\widehat{V}(\widehat{T}_{xSS}) = \frac{1}{1-\sum_{k\in s} a_k^2}\sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l x_l/\hat{\pi}(\delta_l))^2.$$

Las covarianzas se estiman como

$$\widehat{cov}(\widehat{T}_{yxSS}, \widehat{T}_{ySS}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{y_k x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l x_l/\hat{\pi}(\delta_l))(\frac{y_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l/\hat{\pi}(\delta_l)),$$

$$\widehat{cov}(\widehat{T}_{yxSS}, \widehat{T}_{xSS}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{y_k x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l x_l/\hat{\pi}(\delta_l))(\frac{x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l x_l/\hat{\pi}(\delta_l)),$$

y

$$\widehat{cov}(\widehat{T}_{ySS}, \widehat{T}_{xSS}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{y_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l/\hat{\pi}(\delta_l))(\frac{x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l x_l/\hat{\pi}(\delta_l)),$$

donde $a_k = (1-\hat{\pi}(\delta_k))/\sum_{l\in s}(1-\hat{\pi}(\delta_l))$.

Como los estimadores de coeficientes de correlación son estimadores de razón, los estimadores de sus varianzas pueden obtenerse fácilmente utilizando la linealización de Taylor (ver, por ejemplo, [121]).

4.4.5. Estudio de simulación

En esta sección se lleva a cabo un estudio de simulación para estudiar el rendimiento del desempeño de los estimadores propuestos bajo diferentes escenarios. El principal factor de interés será la estimación de la covarianza poblacional y la correlación entre covariables continuas.

El tamaño de la población simulada es $N = 10000$. Se genera aleatoriamente una sociomatriz Z de conexión de red como la definida en la sección 2.2. El resultado z_{ij} determinará el grado, ya que $\delta_j = \sum_i z_{ij}$. Se seleccionan 10 semillas al azar de la red con probabilidad proporcional a su grado, con 3 cupones máximos emitidos para cada participante.

Los valores de la variable de interés y se generan a partir de una distribución normal $y_j \sim N(5000, 500)$, para $j = 1, \ldots, 5000$. Se generan tres variables auxiliares a partir de los valores de y, que son: $x_1 = (y - e_1)/0{,}5$ con $e_1 \sim N(500, 500)$, $x_2 = (y - e_2)/0{,}5$ con $e_2 \sim N(500, 700)$ y $x_3 = (y - e_3)/0{,}5$, donde $e_3 \sim N(500, 300)$. Los coeficientes de correlación resultantes son $\rho = 0{,}7007$ para x_1, $\rho = 0{,}571$ para x_2 y $\rho = 0{,}8579$ para x_3, respectivamente. El tamaño de la muestra es $n = 500$ y las muestras se seleccionan mediante un muestreo aleatorio simple sin reemplazo, al igual que RDS se suele realizar en la práctica.

Para cada modelo de regresión, se calculan los dos estimadores propuestos de la covarianza poblacional S_{yx} y el coeficiente de correlación ρ. Se considera el porcentaje de sesgo relativo

$$rb\,\% = E_{MC}(\widehat{\theta} - \theta)/\theta * 100,$$

y el porcentaje de error cuadrático medio relativo

$$rmse\,\% = E_{MC}[(\widehat{\theta} - \theta)^2]/\theta^2 * 100,$$

para cada estimador $\widehat{S}_{yx}$ y $\widehat{\rho}$. Los resultados de la simulación se basan en $B = 1000$ muestras y E_{MC} denota el promedio de réplicas de Monte Carlo.

Tabla 4.1: Sesgo porcentual relativo ($rb\,\%$) y error cuadrático medio relativo en tanto por ciento ($rmse\,\%$) para estimar S_{yx} con estimadores $\widehat{S}_{yx,RDS\ II}$ y $\widehat{S}_{yx,RDS\ SS}$ en los tres escenarios.

Estimador	$\widehat{S}_{yx,RDS\ II}$		$\widehat{S}_{yx,RDS\ SS}$	
	$rb\,\%$	$rmse\,\%$	$rb\,\%$	$rmse\,\%$
Escenario 1	1.4953	0.8158	1.5055	0.8065
Escenario 2	1.7857	1.0906	1.7897	1.0782
Escenario 3	1.2745	0.6516	1.2924	0.6447

Tabla 4.2: Sesgo porcentual relativo ($rb\,\%$) y error cuadrático medio relativo en tanto por ciento ($rmse\,\%$) para estimar el coeficiente de correlación ρ con los estimadores $\widehat{\rho}_{RDS\ II}$ y $\widehat{\rho}_{RDS\ SS}$ en los tres escenarios.

Estimador	$\widehat{\rho}_{RDS\ II}$		$\widehat{\rho}_{RDS\ SS}$	
	$rb\,\%$	$rmse\,\%$	$rb\,\%$	$rmse\,\%$
Escenario 1	0.4738	0.1262	0.4786	0.1245
Escenario 2	0.8656	0.3347	0.8628	0.3304
Escenario 3	0.1889	0.0244	0.2003	0.0242

Los estimadores de la covarianza son aproximadamente insesgados, ya que los sesgos relativos rondan el 1 % para todos los escenarios considerados, con sesgos aún menores para las estimaciones del coeficiente

de correlación, todos ellos menores al 1 % como se muestran en las Tablas 4.1 y 4.2. Pequeños valores de eficiencia relativa para la estimación de los parámetros con resultados bastante similares obtenidos con ambos estimadores, lo que indica que son efectivos en la estimación de estos parámetros no lineales.

4.4.6. Aplicación en la encuesta RDS de minorías étnicas en Ecuador

En esta sección, los estimadores propuestos se aplican a una encuesta real que estudia la discriminación y subrepresentación de jóvenes indígenas, monubios y afroecuatorianos en Ecuador que se describe en el apartado 3.3 de esta tesis. La metodología RDS se aplicó a una población de jóvenes (18 a 29 años) indígenas, monubios y afroecuatorianos residentes en la ciudad de Riobamba (Ecuador). Históricamente han sufrido exclusión y subrepresentación y, por lo tanto, este clúster carece de un marco muestral confiable [98, 99, 100, 101]. Se reclutó a 814 personas en seis oleadas y se les preguntó sobre sus antecedentes sociales y económicos y sus condiciones de vida mediante un sistema dual de incentivos para motivar la contratación. El ingreso informado del hogar es la variable de interés y la edad del entrevistado es la covariable.

Tabla 4.3: Estimaciones de b (coeficiente de regresión) y S_{xy} para el ejemplo étnico.

$\widehat{b}_{RDS\ II}$	$\widehat{b}_{RDS\ SS}$	$\widehat{S}_{yx,RDS\ II}$	$\widehat{S}_{yx,RDS\ SS}$
24.2108	24.8217	310.2961	322.5241

Tabla 4.4: Sesgo porcentual relativo ($rb\,\%$) y error cuadrático medio relativo en tanto por ciento ($rmse\,\%$) para estimar S_{yx} con estimadores $\widehat{S}_{yx,RDS\ II}$ y $\widehat{S}_{yx,RDS\ SS}$ para el ejemplo étnico.

$\widehat{S}_{yx,RDS\ II}$		$\widehat{S}_{yx,RDS\ SS}$	
$rb\,\%$	$rmse\,\%$	$rb\,\%$	$rmse\,\%$
-5.6710	0.8551	-7.1474	0.5216

Tabla 4.5: Sesgo porcentual relativo ($rb\,\%$) y error cuadrático medio relativo en tanto por ciento ($rmse\,\%$) para estimar el coeficiente de correlación ρ con los estimadores $\widehat{\rho}_{RDS\ II}$ y $\widehat{\rho}_{RDS\ SS}$ para el ejemplo étnico.

$\widehat{\rho}_{RDS\ II}$		$\widehat{\rho}_{RDS\ SS}$	
$rb\,\%$	$rmse\,\%$	$rb\,\%$	$rmse\,\%$
6.1371	0.4460	4.9065	0.2407

Los valores estimados del coeficiente de regresión y de covarianza se presentan en la Tabla 4.3. Se obtiene un buen desempeño general de los dos estimadores propuestos para la covarianza y el coeficiente de correlación, con un sesgo de aproximadamente 5 % y valores pequeños similares del error cuadrático medio relativo $rmse$, como se muestran en las Tablas 4.4 y 4.5.

Capítulo 5

Conclusiones

La motivación de esta tesis fue en primer lugar, poner en práctica una encuesta no probabilística mediante la metodología RDS, para recoger información sobre poblaciones indígenas y otras minorías étnicas. En segundo lugar, fue formular expresiones de estimadores de covarianza y coeficientes de correlación en el contexto del muestreo de poblaciones difíciles de alcanzar. El cumplimiento de estos objetivos motivó el estudio de las encuestas RDS de minorías étnicas y la revisión de la literatura sobre análisis de regresión en el muestreo dirigido por los participantes.

El muestreo dirigido por los participantes combina un método eficiente de muestreo de referencia en cadena con un método de análisis estadístico que corrige el hecho de que los datos se recopilan de forma no aleatoria, para proporcionar estimaciones de población no sesgadas. Se ha utilizado ampliamente en los campos de la Salud pública y la Sociología para estudiar poblaciones ocultas como las que están en riesgo de contraer VIH, la comunidad LGBTI, Minorías étnicas, inmigrantes y niños de la calle.

En el contexto de poblaciones ocultas y/o difíciles de alcanzar, las cuales generalmente les hace falta una muestra aleatoria representativa y/o no tienen un marco de muestreo, las principales ventajas de RDS son: i) el mejoramiento de la eficiencia del reclutamiento de los encuestados, ii) la reducción del estigma o preocupación de la participación en el estudio, iii) la posibilidad de combinar la metodología RDS con internet en teléfonos móviles, tablets, páginas web y redes sociales como estrategia para la recolección de datos, iv) la rentabilidad, debido a que la implementación de RDS generalmente es más económico que el muestreo tradicional y v) la estimación asintóticamente no sesgados para las características de la

población bajo ciertos supuestos.

En contraste las principales desventajas de RDS son: i) la selección de muestras basadas en la red social, que dependen en gran medida de las conexiones de redes sociales entre los miembros de la población y ii) las estimaciones basadas en supuestos rigurosos que, en la práctica, casi ningún supuesto de los estimadores RDS puede cumplirse. Debido a esto, se recomienda que los usuarios de RDS investiguen a fondo la violación de las suposiciones, hagan ajustes si es posible e interpreten con precaución los resultados de las muestras de RDS.

5.1. Contribuciones

Se ha demostrado que, RDS es un método eficaz para analizar la estructura de las redes sociales de minorías étnicas y se ha implementado con éxito como método de muestreo en la web en Ecuador, un país en el que las poblaciones minoritarias como indígenas, montubios y afroecuatorianos están muy estigmatizadas y subrepresentadas en las encuestas oficiales. En este contexto, se utilizó con éxito RDS en la web para muestrear y encuestar a 814 jóvenes urbanos sobre una serie de temas delicados. La muestra obtenida es étnica, demográfica y socioeconómicamente diversa y suficientemente grande como para producir estimaciones sobre la población. Al comparar las estadísticas oficiales publicadas con nuestras estimaciones, se ha demostrado que existen diferencias en algunas características socioeconómicas entre quienes se autoidentifican como parte de las minorías étnicas de estudio y quienes son reacios a hacerlo. Por otro lado, de nuestra experiencia RDS en la web implicará un coste menor que un estudio RDS estándar. Por lo tanto, RDS en la web proporciona potencialmente un medio útil para llegar a poblaciones ocultas conectadas en la web.

Dentro del desarrollo de métodos de regresión para RDS, se ha propuesto un nuevo método de estimación del peso de la muestra para datos continuos. El enfoque propuesto es más apropiado para situaciones en las que la homofilia es pequeña. Como ilustración de la aplicabilidad del método propuesto, se ha realizado un estudio de simulación y una aplicación a un ejemplo étnico (en la muestra RDS web en Ecuador). Sin embargo, el enfoque de nuestro trabajo ha sido proponer un método para estimar parámetros no lineales con nuevos pesos muestrales. Se ha derivado expresiones de las varianzas y también se ha demostrado que los estimadores propuestos tienen propiedades deseables. El estudio de simulación no muestra diferencias significativas en términos de sesgo o error cuadrático medio entre los

dos estimadores propuestos. Además, la complejidad de cálculo de los dos estimadores es similar. Por tanto, no hay ninguna razón objetiva para preferir una sobre la otra.

Tomados en conjunto, los resultados sobre la dependencia entre variables continuas presentados en esta tesis, se suman a la creciente literatura sobre muestreo dirigido por los participantes, lo que permite a los investigadores obtener mejor información sobre las poblaciones ocultas de interés.

5.2. Implicaciones de la tesis

RDS en la web proporciona a las personas estigmatizadas un acceso rápido y fácil al estudio, con una exposición mínima de su identidad. La implementación exitosa del estudio de minorías étnicas de indígenas, montubios y afroecuatorianos demuestra que es posible reclutar grupos de difícil acceso a través de internet. Debido a que RDS en la web es capaz de proporcionar estimaciones de población y es de fácil acceso, también puede ser útil para estudios de otros tipos de poblaciones en línea cuando hay una falta de marco muestral. Ejemplos de aplicaciones incluyen usuarios de correo electrónico universitario, usuarios registrados en foros web, usuarios de Facebook, etc.

Para los responsables de la formulación de políticas públicas, el hallazgo importante es que la discriminación de los grupos minoritarios sigue estando presente en la actualidad y es necesario diseñar nuevas políticas para erradicar todas las formas de discriminación. Eliminar la discriminación es un objetivo alcanzable y digno, entonces esta debería ser un área prioritaria de investigación y enfoque de políticas. En el mismo sentido, es importante investigaciones futuras sobre el efecto de la discriminación en las condiciones de vida de las minorías étnicas en el contexto de RDS, mediante el modelado de regresión y asociación para datos RDS.

El estudio de indicadores económicos, como las medidas de pobreza, es cada vez más relevante para la sociedad y para los responsables políticos. Por lo tanto, cuando nos interesa estudiar medidas de pobreza en el contexto de poblaciones difíciles de alcanzar y/o ocultas, por ejemplo, las minorías étnicas, es importante desarrollar estimadores de la función de distribución y cuantiles en un entorno de muestreo no probabilístico. Este tema debería ser parte de la investigación futura.

Por otra parte, si bien se considera que el nuevo método de estimación del peso de la muestra para datos continuos es un enfoque novedoso para los datos RDS continuos, la contabilidad de la agrupación sigue siendo una cuestión abierta. Es posible extender esta metodología a la adaptación a clústeres, como

parte de futuras investigaciones.

5.3. Limitaciones

RDS en la web está principalmente limitado por el requisito de que las personas de la población objetivo tengan acceso frecuente a internet (sesgo de cobertura), es decir, las personas que no están conectadas en la web no pueden ser reclutadas o reclutar a otras en sus redes. Además, en poblaciones donde el uso de internet es muy variable, el período de muestreo debe permanecer abierto el tiempo suficiente para que los usuarios de internet revisen sus buzones de entrada y respondan. También, es difícil identificar si un participante de RDS en la web es realmente parte de la población de minoría étnica. Además, como sucede en la mayoría de los estudios empíricos de RDS, nuestra muestra RDS web en Ecuador sufre varias violaciones de los supuestos de RDS. Sin embargo, como no es posible evaluar el nivel de todas las posibles violaciones, la interpretación de las estimaciones de población de la muestra de RDS en la web debe estar condicionado a estas incertidumbres. Por otro lado, no está claro cómo las respuestas de los encuestados podrían verse influido por la respuesta socialmente deseable en encuestas que utilizan métodos de muestreo RDS en la web, por lo tanto, es necesario investigación futura sobre este tema.

Finalmente, sobre el sesgo de cobertura, es importante explorar el empleo de métodos de marcos múltiples con RDS en la web. Las encuestas con marcos múltiples pueden mejorar notablemente la eficiencia de un conjunto de datos, especialmente cuando se muestrea una población difícil de alcanzar y/o oculta, es decir, un subgrupo de interés que comprende solamente una pequeña parte del total de la población. Por lo tanto, es importante que la investigación futura trate sobre este tema.

Apéndice

.1. Estimaciones RDS I, RDS II y RDS SS en la encuesta RDS de minorías étnicas en Ecuador

Tabla 1: Estimaciones RDS I, RDS II y RDS SS para todas las variables de la encuesta.

Variable	**Categoría**	**Estimación RDS**		**Intervalo Confiianza al 95 %**		**Efecto Diseño**	**SD**	**n**
Sexo	Hombre	RDS I	0.4268	0.3896	0.4640	1.2865	0.0190	
		RDS II	0.4254	0.3460	0.5048	5.8653	0.0405	364
		RDS SS	0.4261	0.3470	0.5051	5.8066	0.0403	
	Mujer	RDS I	0.5732	0.5360	0.6104	1.2865	0.0190	
		RDS II	0.5746	0.4952	0.6540	5.8653	0.0405	450
		RDS SS	0.5739	0.4949	0.6530	5.8066	0.0403	
	Edad	RDS I	21.7910	21.5480	22.0340	1.3243	0.1241	
		RDS II	21.8060	21.2930	22.3180	5.7834	0.2615	814
		RDS SS	21.8130	21.3020	22.3240	5.7469	0.2607	
Estado civil	Casado-Unión libre	RDS I	0.1783	0.1523	0.2043	1.0471	0.0133	
		RDS II	0.1820	0.1321	0.2319	3.7991	0.0254	163
		RDS SS	0.1827	0.1330	0.2323	3.7594	0.0253	
	Divorciado-separado	RDS I	0.0040	0.0008	0.0072	0.5995	0.0017	
		RDS II	0.0036	0.0015	0.0057	0.2821	0.0011	4
		RDS SS	0.0037	0.0015	0.0058	0.2806	0.0011	
	Soltero	RDS I	0.8170	0.7908	0.8432	1.0439	0.0134	
		RDS II	0.8138	0.7638	0.8639	3.7616	0.0255	646
		RDS SS	0.8131	0.7632	0.8630	3.7236	0.0254	
	Viudo	RDS I	0.0007	0.0000	0.0016	0.2444	0.0004	
		RDS II	0.0005	0.0003	0.0008	0.0271	0.0001	1
		RDS SS	0.0006	0.0003	0.0008	0.0316	0.0001	
Autoidentificación étnica	Afroecuatoriano	RDS I	0.0456	0.0349	0.0564	0.6019	0.0055	
		RDS II	0.0502	0.0322	0.0682	1.5400	0.0092	44
		RDS SS	0.0504	0.0324	0.0683	1.5320	0.0092	
	Indígena	RDS I	0.9250	0.9073	0.9428	1.0331	0.0091	
		RDS II	0.9197	0.8798	0.9596	4.8974	0.0203	749
		RDS SS	0.9197	0.8804	0.9590	4.7620	0.0201	
	Montubio	RDS I	0.0294	0.0148	0.0439	1.6982	0.0074	
		RDS II	0.0301	0.0000	0.0665	10.3560	0.0186	21
		RDS SS	0.0299	0.0000	0.0658	10.0820	0.0183	
Cantón	Otro	RDS I	0.5549	0.5151	0.5947	1.4589	0.0203	
		RDS II	0.5502	0.4690	0.6314	6.0572	0.0414	475
		RDS SS	0.5514	0.4699	0.6328	6.0943	0.0415	
	Riobamba	RDS I	0.4451	0.4053	0.4849	1.4589	0.0203	
		RDS II	0.4498	0.3686	0.5310	6.0572	0.0414	339
		RDS SS	0.4486	0.3672	0.5301	6.0943	0.0415	

Tabla 1: *Cont.*

Variable	**Categoría**	**Estimación RDS**		**Intervalo Confiianza al 95 %**		**Efecto Diseño**	**SD**	**n**
Vestimenta	No	RDS I	0.4417	0.4025	0.4809	1.4188	0.0200	
		RDS II	0.4407	0.3587	0.5227	6.2000	0.0418	340
		RDS SS	0.4401	0.3583	0.5219	6.1688	0.0417	
	Si	RDS I	0.5583	0.5191	0.5975	1.4188	0.0200	
		RDS II	0.5593	0.4773	0.6413	6.2000	0.0418	474
		RDS SS	0.5599	0.4781	0.6417	6.1688	0.0417	
Instrucción	Secundaria o menos	RDS I	0.8040	0.7795	0.8284	0.8639	0.0125	
		RDS II	0.7999	0.7559	0.8438	2.7416	0.0224	624
		RDS SS	0.7985	0.7546	0.8424	2.7214	0.0224	
	Al menos la universidad	RDS I	0.1960	0.1716	0.2205	0.8639	0.0125	
		RDS II	0.2001	0.1562	0.2441	2.7416	0.0224	190
		RDS SS	0.2015	0.1576	0.2454	2.7214	0.0224	
Instrucción madre	Secundaria o menos	RDS I	0.8937	0.8772	0.9102	0.6537	0.0084	
		RDS II	0.8897	0.8367	0.9428	6.5288	0.0271	712
		RDS SS	0.8893	0.8375	0.9411	6.2066	0.0264	
	Al menos la universidad	RDS I	0.1063	0.0898	0.1228	0.6537	0.0084	
		RDS II	0.1103	0.0572	0.1633	6.5288	0.0271	102
		RDS SS	0.1107	0.0589	0.1625	6.2066	0.0264	
Instrucción padre	Secundaria o menos	RDS I	0.9426	0.9325	0.9527	0.4280	0.0051	
		RDS II	0.9417	0.9270	0.9564	0.8974	0.0075	754
		RDS SS	0.9412	0.9264	0.9560	0.8955	0.0075	
	Al menos la universidad	RDS I	0.0574	0.0473	0.0675	0.4280	0.0051	
		RDS II	0.0583	0.0436	0.0730	0.8974	0.0075	60
		RDS SS	0.0588	0.0440	0.0736	0.8955	0.0075	
	Número cuartos	RDS I	2.7550	2.6792	2.8308	1.0495	0.0387	
		RDS II	2.8098	2.6223	2.9974	5.9969	0.0957	814
		RDS SS	2.8083	2.6214	2.9952	5.9538	0.0954	
	Número personas	RDS I	4.9082	4.7044	5.1121	1.9424	0.1040	
		RDS II	4.6912	4.4334	4.9490	3.8184	0.1315	814
		RDS SS	4.6901	4.4354	4.9448	3.7184	0.1300	

Tabla 1: *Cont.*

Variable	**Categoría**	**Estimación RDS**		**Intervalo Confiianza al 95 %**		**Efecto Diseño**	**SD**	**n**
Servicio agua	Pozo-grieta, granel u otro	RDS I	0.0802	0.0600	0.1004	1.2623	0.0103	
		RDS II	0.0793	0.0538	0.1049	2.0368	0.0131	66
		RDS SS	0.0793	0.0539	0.1047	2.0048	0.0129	
	Tubería	RDS I	0.9198	0.8996	0.9400	1.2623	0.0103	
		RDS II	0.9207	0.8951	0.9462	2.0368	0.0131	748
		RDS SS	0.9207	0.8953	0.9461	2.0048	0.0129	
Servicio energía	Compañia pública	RDS I	0.9531	0.9381	0.9681	1.1484	0.0077	
		RDS II	0.9525	0.9345	0.9705	1.6318	0.0092	774
		RDS SS	0.9524	0.9345	0.9704	1.6166	0.0092	
	Generador, vela u otro	RDS I	0.0469	0.0319	0.0619	1.1484	0.0077	
		RDS II	0.0475	0.0295	0.0655	1.6318	0.0092	40
		RDS SS	0.0476	0.0296	0.0655	1.6166	0.0092	
Discapacidad	No	RDS I	0.8809	0.8514	0.9105	1.8925	0.0151	
		RDS II	0.8823	0.8243	0.9403	7.3672	0.0296	728
		RDS SS	0.8827	0.8250	0.9405	7.3228	0.0295	
	Si	RDS I	0.1191	0.0895	0.1486	1.8925	0.0151	
		RDS II	0.1177	0.0597	0.1757	7.3672	0.0296	86
		RDS SS	0.1173	0.0595	0.1750	7.3228	0.0295	
Visita	No	RDS I	1.0000	-	-	-	-	
		RDS II	0.6520	0.4504	0.8535	4.0361	0.1028	56
		RDS SS	0.6520	0.4507	0.8533	4.0270	0.1027	
	Si	RDS I	0.0000	-	-	-	-	
		RDS II	0.3480	0.1465	0.5496	4.0361	0.1028	30
		RDS SS	0.3480	0.1467	0.5493	4.0270	0.1027	
Lenguaje	Español	RDS I	0.7578	0.7266	0.7890	1.2048	0.0159	
		RDS II	0.7589	0.6914	0.8264	5.6623	0.0344	608
		RDS SS	0.7585	0.6918	0.8253	5.5309	0.0341	
	Indígena	RDS I	0.2422	0.2110	0.2734	1.2048	0.0159	
		RDS II	0.2411	0.1736	0.3086	5.6623	0.0344	206
		RDS SS	0.2415	0.1747	0.3082	5.5309	0.0341	

Tabla 1: *Cont.*

Variable	**Categoría**	**Estimación RDS**		**Intervalo Confiianza al 95 %**		**Efecto Diseño**	**SD**	**n**
Lenguaje padres	Español	RDS I	0.2100	0.1766	0.2433	1.5252	0.0170	
		RDS II	0.2079	0.1455	0.2703	5.3802	0.0318	156
		RDS SS	0.2074	0.1457	0.2690	5.2522	0.0314	
	Extranjero	RDS I	0.0023	0.0000	0.0079	3.1621	0.0029	
		RDS II	0.0018	0.0003	0.0032	0.2714	0.0007	1
		RDS SS	0.0018	0.0003	0.0032	0.2834	0.0008	
	Extranjero-Español	RDS I	0.0020	0.0000	0.0118	10.7650	0.0050	
		RDS II	0.0018	0.0000	0.0057	2.0257	0.0020	1
		RDS SS	0.0018	0.0000	0.0058	2.0728	0.0020	
	Indígena	RDS I	0.1034	0.0788	0.1281	1.4900	0.0126	
		RDS II	0.1031	0.0605	0.1457	4.4593	0.0217	81
		RDS SS	0.1031	0.0611	0.1450	4.3376	0.0214	
	Indígena-Español	RDS I	0.6823	0.6442	0.7205	1.5268	0.0195	
		RDS II	0.6855	0.6150	0.7560	5.2381	0.0360	575
		RDS SS	0.6861	0.6165	0.7557	5.1144	0.0355	
Inscripción	Si	RDS I	0.5480	0.5092	0.5868	1.3825	0.0198	
		RDS II	0.5398	0.4593	0.6203	5.9341	0.0411	452
		RDS SS	0.5405	0.4609	0.6201	5.8004	0.0406	
	No	RDS I	0.4520	0.4132	0.4908	1.3825	0.0198	
		RDS II	0.4602	0.3797	0.5407	5.9341	0.0411	362
		RDS SS	0.4595	0.3799	0.5391	5.8004	0.0406	
Razón para no inscribirse	Otro	RDS I	0.7702	0.6904	0.8500	3.5045	0.0407	
		RDS II	0.8079	0.6800	0.9359	10.2810	0.0653	306
		RDS SS	0.8088	0.6803	0.9373	10.4040	0.0656	
	Finalización de estudios	RDS I	0.2298	0.1500	0.3096	3.5045	0.0407	
		RDS II	0.1921	0.0641	0.3200	10.2810	0.0653	57
		RDS SS	0.1912	0.0627	0.3197	10.4040	0.0656	
Tipo de escuela	Estado	RDS I	0.8921	0.8627	0.9216	1.1049	0.0150	
		RDS II	0.8981	0.8539	0.9423	2.6084	0.0225	401
		RDS SS	0.8977	0.8538	0.9417	2.5703	0.0224	
	Privada	RDS I	0.1079	0.0784	0.1373	1.1049	0.0150	
		RDS II	0.1019	0.0577	0.1461	2.6084	0.0225	51
		RDS SS	0.1023	0.0583	0.1462	2.5703	0.0224	

Tabla 1: *Cont.*

Variable	Categoría	Estimación RDS		Intervalo Confiianza al 95 %		Efecto Diseño	SD	n
		RDS I	301.81	281.49	322.13	0.4440	10.367	
	Salario	RDS II	294.60	258.43	330.77	1.7861	18.453	439
		RDS SS	295.50	259.60	331.39	1.7204	18.315	
		RDS I	0.8972	0.8768	0.9177	0.5552	0.0104	
	Español	RDS II	0.8740	0.8403	0.9078	1.2645	0.0172	386
		RDS SS	0.8733	0.8393	0.9072	1.2735	0.0173	
		RDS I	0.0118	0.0040	0.0195	0.6357	0.0040	
	Extranjero	RDS II	0.0125	0.0013	0.0237	1.2494	0.0057	7
		RDS SS	0.0126	0.0013	0.0239	1.2486	0.0057	
		RDS I	0.0616	0.0468	0.0765	0.4672	0.0076	
	Extranjero-Español	RDS II	0.0742	0.0503	0.0982	1.0181	0.0122	39
Idioma		RDS SS	0.0747	0.0507	0.0986	1.0184	0.0122	
clase	Extranjero-	RDS I	0.0146	0.0079	0.0213	0.3825	0.0034	
	Español-	RDS II	0.0173	0.0092	0.0254	0.4697	0.0041	10
	Indígena	RDS SS	0.0175	0.0093	0.0257	0.4823	0.0042	
		RDS I	0.0036	0.0000	0.0076	0.5446	0.0020	
	Indígena	RDS II	0.0043	0.0000	0.0098	0.8708	0.0028	2
		RDS SS	0.0043	0.0000	0.0098	0.8739	0.0028	
		RDS I	0.0112	0.0073	0.0151	0.1696	0.0020	
	Indígena-Español	RDS II	0.0177	0.0110	0.0245	0.3205	0.0034	8
		RDS SS	0.0177	0.0109	0.0246	0.3277	0.0035	
		RDS I	0.5046	0.4648	0.5444	1.4400	0.0203	
	No	RDS II	0.5021	0.4224	0.5818	5.7754	0.0407	375
		RDS SS	0.5009	0.4216	0.5801	5.7114	0.0404	
Trabajo		RDS I	0.4954	0.4556	0.5352	1.4400	0.0203	
	Si	RDS II	0.4979	0.4182	0.5776	5.7754	0.0407	439
		RDS SS	0.4991	0.4199	0.5784	5.7114	0.0404	
		RDS I	0.1487	0.1194	0.1781	0.8064	0.0150	
	Asegurado	RDS II	0.1500	0.0973	0.2027	2.5840	0.0269	71
Seguridad		RDS SS	0.1505	0.0983	0.2026	2.5221	0.0266	
social		RDS I	0.8513	0.8219	0.8806	0.8064	0.0150	
	Sin seguro	RDS II	0.8500	0.7973	0.9027	2.5840	0.0269	368
		RDS SS	0.8495	0.7974	0.9017	2.5221	0.0266	

Tabla 1: *Cont.*

Variable	**Categoría**	**Estimación RDS**		**Intervalo Confiianza al 95 %**		**Efecto Diseño**	**SD**	**n**
Ocupación	Directores de administración	RDS I	0.0190	0.0127	0.0253	0.2513	0.0032	
		RDS II	0.0399	0.0000	0.0838	5.9572	0.0224	15
		RDS SS	0.0397	0.0000	0.0826	5.7393	0.0219	
	Empleados administrativos	RDS I	0.0628	0.0457	0.0798	0.5866	0.0087	
		RDS II	0.0678	0.0353	0.1003	1.9862	0.0166	33
		RDS SS	0.0680	0.0359	0.1002	1.9356	0.0164	
	Operadores de instalaciones y maquinaria	RDS I	0.0250	0.0098	0.0401	1.1213	0.0077	
		RDS II	0.0171	0.0071	0.0270	0.7012	0.0051	8
		RDS SS	0.0171	0.0073	0.0269	0.6763	0.0050	
	Operadores y artesanos	RDS I	0.1116	0.0722	0.1510	1.8564	0.0201	
	Operadores y artesanos	RDS II	0.0890	0.0394	0.1387	3.6012	0.0253	40
		RDS SS	0.0891	0.0404	0.1379	3.4739	0.0249	
	Científicos e intelectuales	RDS I	0.0417	0.0220	0.0613	1.1491	0.0100	
		RDS II	0.0397	0.0084	0.0710	3.0509	0.0160	15
		RDS SS	0.0395	0.0088	0.0702	2.9523	0.0157	
	Técnicos de nivel medio	RDS I	0.0342	0.0120	0.0564	1.7716	0.0113	
		RDS II	0.0298	0.0136	0.0460	1.0788	0.0083	12
		RDS SS	0.0296	0.0135	0.0457	1.0716	0.0082	
	Agrícola y pesquero	RDS I	0.0141	0.0102	0.0179	0.1260	0.0020	
		RDS II	0.0166	0.0108	0.0223	0.2389	0.0029	13
		RDS SS	0.0171	0.0112	0.0229	0.2449	0.0030	
	Trabajadores de servicios y comercio	RDS I	0.1896	0.1498	0.2295	1.2255	0.0203	
								98
		RDS II	0.1961	0.1257	0.2666	3.7358	0.0359	
		RDS SS	0.1971	0.1273	0.2669	3.6485	0.0356	
	Trabajadores no calificados	RDS I	0.5020	0.4481	0.5560	1.3822	0.0275	
		RDS II	0.5040	0.4005	0.6075	5.0865	0.0528	205
		RDS SS	0.5027	0.3996	0.6059	5.0480	0.0526	
Tierra	No	RDS I	0.6374	0.6001	0.6747	1.3689	0.0190	
		RDS II	0.6414	0.5675	0.7153	5.4001	0.0377	512
		RDS SS	0.6408	0.5681	0.7135	5.2201	0.0371	
	Si	RDS I	0.3626	0.3253	0.3999	1.3689	0.0190	
		RDS II	0.3586	0.2847	0.4325	5.4001	0.0377	302
		RDS SS	0.3592	0.2865	0.4319	5.2201	0.0371	

Tabla 1: *Cont.*

Variable	Categoría	Estimación RDS		Intervalo Confiianza al 95 %		Efecto Diseño	SD	n
Cosecha	No	RDS	I	0.2524	0.1989	0.3060	1.2255	0.0273	
		RDS	II	0.2504	0.1139	0.3868	8.0024	0.0696	76
		RDS	SS	0.2505	0.1154	0.3855	7.8356	0.0689	
	Si	RDS	I	0.7476	0.6940	0.8011	1.2255	0.0273	
		RDS	II	0.7496	0.6132	0.8861	8.0024	0.0696	226
		RDS	SS	0.7495	0.6145	0.8846	7.8356	0.0689	
Relación laboral	Asalariado	RDS	I	0.3594	0.3021	0.4167	1.6911	0.0292	
		RDS	II	0.3420	0.2357	0.4483	5.9597	0.0542	138
		RDS	SS	0.3411	0.2364	0.4458	5.7840	0.0534	
	Cuenta propia	RDS	I	0.3754	0.3256	0.4252	1.2555	0.0254	
		RDS	II	0.3925	0.2918	0.4933	5.0492	0.0514	183
		RDS	SS	0.3935	0.2945	0.4925	4.8738	0.0505	
	Empleado doméstico	RDS	I	0.1275	0.0894	0.1657	1.5496	0.0195	
		RDS	II	0.1237	0.0681	0.1794	3.3866	0.0284	51
		RDS	SS	0.1234	0.0683	0.1784	3.3275	0.0281	
	Empleador o socio activo	RDS	I	0.0351	0.0189	0.0513	0.9242	0.0083	
		RDS	II	0.0331	0.0162	0.0500	1.0615	0.0086	16
		RDS	SS	0.0332	0.0163	0.0501	1.0588	0.0086	
	Trabajador familiar no remunerado	RDS	I	0.1025	0.0758	0.1293	0.9216	0.0136	
		RDS	II	0.1086	0.0639	0.1534	2.4549	0.0228	51
		RDS	SS	0.1089	0.0644	0.1533	2.4155	0.0227	
Animales	No	RDS	I	0.2879	0.2190	0.3567	1.8633	0.0351	
		RDS	II	0.2569	0.1231	0.3907	7.5658	0.0683	73
		RDS	SS	0.2566	0.1234	0.3897	7.4986	0.0679	
	Si	RDS	I	0.7121	0.6433	0.7810	1.8633	0.0351	
		RDS	II	0.7431	0.6093	0.8769	7.5658	0.0683	229
		RDS	SS	0.7434	0.6103	0.8766	7.4986	0.0679	
Pobreza	Pobreza extrema	RDS	I	0.0902	0.0677	0.1126	1.3843	0.0114	
		RDS	II	0.0938	0.0509	0.1366	4.8772	0.0219	71
		RDS	SS	0.0935	0.0506	0.1364	4.8907	0.0219	
	Sin pobreza extrema	RDS	I	0.9098	0.8874	0.9323	1.3843	0.0114	
		RDS	II	0.9062	0.8634	0.9491	4.8772	0.0219	737
		RDS	SS	0.9065	0.8636	0.9494	4.8907	0.0219	

Tabla 1: *Cont.*

Variable	**Categoría**	**Estimación RDS**		**Intervalo Confiianza al 95 %**		**Efecto Diseño**	**SD**	**n**
Víctima	Frecuentemente	RDS	I	0.0925	0.0748	0.1102	0.8476	0.0090	
		RDS	II	0.0932	0.0502	0.1362	4.9735	0.0219	87
		RDS	SS	0.0937	0.0507	0.1367	4.9493	0.0219	
	Ocasionalmente	RDS	I	0.9075	0.8898	0.9252	0.8476	0.0090	
		RDS	II	0.9068	0.8638	0.9498	4.9735	0.0219	727
		RDS	SS	0.9063	0.8633	0.9493	4.9493	0.0219	
	Satisfacción general	RDS	I	8.5455	8.4242	8.6669	1.6210	0.0619	
		RDS	II	8.5468	8.2907	8.8028	6.9519	0.1306	814
		RDS	SS	8.5506	8.2982	8.8029	6.7684	0.1288	

Bibliografía

[1] Sudman, S.; Kalton, G. New Developments in the Sampling of Special Populations. *Annual Review of Sociology.* **1986**, *1*, 401–429.

[2] Sydor, A. Conducting research into hidden or hard-to-reach populations. *Nurse researcher.* **2013**, *20(3)*, 33–37.

[3] Spreen, M.; Zwaagstra, R. Personal network sampling, outdegree analysis and multilevel analysis: Introducing the network concept in studies of hidden populations. *International Sociology.* **1994**, *9(4)*, 475–491.

[4] Heckathorn, D. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems.* **1997**, *44(2)*, 174–199.

[5] Salganik, M.J.; Heckathorn, D.D. 5. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology.* **2004**, *34*, 193–240, doi:10.1111/j.0081-1750.2004.00152.x.

[6] Volz, E.; Heckathorn, D. Probability based estimation theory for respondent driven sampling. *Journal of official statistics.* **2008**, *14(1)*, 79–97.

[7] Kan, M.; Garfinkel, D.; Samoylova, O.; Gray, R.; Little, K. Social network methods for HIV case-finding among people who inject drugs in Tajikistan. *Journal of the International AIDS Society.* **2018**, *21(S5)*, 57–64.

[8] Sypsa, V.; Psichogiou, M.; Paraskevis, D.; Rapid decline in HIV incidence among persons who inject drugs during a fast-track combination prevention program after an HIV outbreak in Athens. *The Journal of infectious diseases.* **2017**, *215(10)*, 1496–1505.

[9] Card, K.G.; Lachowsky, N.J.; Cui, Z. Exploring the role of sex-seeking apps and websites in the social and sexual lives of gay, bisexual and other men who have sex with men: a cross-sectional study. *Sex Health.* **2017**, *14(3)*, 229–237.

[10] Rotondi, M.A.; O'Campo, P.; O'Brien, K. Our Health Counts Toronto: using respondent-driven sampling to unmask census undercounts of an urban indigenous population in Toronto, Canada. *BMJ Open.* **2017**, *7*:e018936, doi:10.1136/bmjopen-2017-018936.

[11] Font, J.; Méndez, M., Eds. *Surveying Ethnic Minorities and Immigrant Populations: Methodological Challenges and Research Strategies.* Amsterdam University Press: Amsterdam, Holland, 2013.

[12] Johnston, L.G.; Thurman, T.R.; Mock, N.; Nano, L.; Carcani, V. Respondent-driven sampling: A new method for studying street children with findings from Albania. *Vulnerable Children and Youth Studies.* **2010**, *5*, 1–11, doi:10.1080/17450120903193923.

[13] Goodman, L. A. Snowball Samplin. *Annals of Mathematical Statist.* **1961**, *32*, 148–170.

[14] Erickson, B. H. Some problems of inference from chain data. *Sociological Methodolog.* **1979**, *10*, 276–30.

[15] Deaux, E.; Callagha, J. Key informant versus self-report estimates of health behavior. *Evaluation Review.* **1985**, *9*, 365–368.

[16] Watters, J.; Patrick, B. Targeted sampling: Options for the study of hidden populations. *Social Problem.* **1989**, *36(4)*, 416–430.

[17] Magnani, R.; Sabin, K.; Saidel, T.; Heckathorn, D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *Aids.* **2005**, *19*, S67–S72.

[18] Heckathorn, D. Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Social problems.* **2002**, *49*, 11–34, doi:10.1525/sp.2002.49.1.11.

[19] Muhib, F.B.; Lin, L.S.; Stueve, A.; Miller, RL.; Ford, W.L.; Johnson, W.D.; Smith, P.J. A venue-based method for sampling hard-to-reach populations. *Public Health Rep.* **2001**, *116(1)*, 216–222.

[20] Killworth, P.; Bernar, R. The reversal small world experiment. *Social Networks.* **1978/79**, *1*, 159–192.

[21] Jonsson J.; Stein, M.; Johansson, G.; Bodin, T.; Strömdahl, S. A performance assessment of web-based respondent driven sampling among workers with precarious employment in Sweden. *PLoS ONE.* **2019**, *14*, doi.org/10.1371/journal.pone.0210183.

[22] Abdesselam, K.; Verdery, A.; Pelude, L.; Dhami, P.; Momoli, F.; Jolly, A.M. The development of respondent-driven sampling (RDS) inference: A systematic review of the population mean and variance estimates. *Drug and alcohol dependence.* **2020**, *206*, doi:10.1016/j.drugalcdep.2019.107702.

[23] Gile, K.J.; Beaudry, I.S.; Handcock, M.S.; Ott, M.Q. Methods for inference from respondent-driven sampling data. *Annual Review of Statistics and Its Application.* **2018**, *5*, 65–93.

[24] Ramirez-Valles, J.; Molina, Y.; Dirkes, J. Stigma towards plwha: the role of internalized homosexual stigma in latino gay/bisexual male and transgender communities. *AIDS Education and Prevention* **2013**, *25*, 179–189.

[25] Johnston, L.R.; O'Bra, H.; Chopra, M.; Mathews, C.; Townsend, L.; Sabin, K.; Tomlinson, M.; Kendall, C. The associations of voluntary counseling and testing acceptance and the perceived likelihood of being HIV-infected among men with multiple sex partners in a South African township. *AIDS and Behavior* **2010**, *14(4)*, 922–931.

[26] Rhodes, S.D.; McCoy, T.P. Condom use among immigrant latino sexual minorities: multilevel analysis after respondent-driven sampling. *AIDS Education and Prevention* **2015**, *27(1)*, 27–43.

[27] Spiller, M.W. Regression modeling of data collected using respondent-driven sampling. Master's Thesis, Cornell University, 2009.

[28] Yauck, M.; Moodie, E.E.; Apelian, H.; Fourmigue, A.; Grace, D.; Hart, T.; Lambert, G.; Cox, J. General Regression Methods for Respondent-Driven Sampling Data. *eprint arXiv:2012.00457.* **2020**.

[29] Beckett, M.; Firestone, M.A.; McKnight, C.D. A cross-sectional analysis of the relationship between diabetes and health access barriers in an urban First Nations population in Canada. *BMJ Open.* **2017**, *8*, e018272.

[30] Schonlau, M.; Liebau, E. Respondent-driven sampling. *The Stata Journal.* **2012**, *12(1)*, 72-93.

[31] Gile, K.; Handcock, M. Respondent-driven sampling: An assessment of current methodology. *Sociological methodology.* **2010**, *40(1)*, 285–327.

[32] Gile, K.J. Inference from Partially-Observed Network Data. PhDdissertation, Department of Statistics, University of Washington, EE.UU, 2008.

[33] Horvitz, D.; Thompson, D.J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association.* **1952**, *42*, 663–685.

[34] Thompson, S.K. *Sampling*, New York: Wiley, 1992.

[35] Gile, K.J.; Johnston, L.G.; Salganik, M.J. Diagnostics for Respondent-driven Sampling *Journal of the Royal Statistical Society.* **2015**, *178(1)*, 241–269, doi:10.1111/rssa.12059.

[36] Hansen, H.; Hurwitz, W. On the Theory of Samplingfrom Finite Populations. *Annals of Mathematical Statistics.* **1943**, *14(4)*, 333–62.

[37] Hastings, W.K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika.* **1970**, *57*, 97–109.

[38] Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. The Monte Carlo Method. *Journal of Chemical Physics.* **1953**, *21*, 21, 1087.

[39] Cochran, W. *Sampling Techniques*, 3d ed. New York: Wiley, 1977.

[40] Brewer, K.R.; Hanif, M. *Sampling with Unequal Probability.* New York: Springer-Verlag, 1983.

[41] Heckathorn, D.D. Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology.* **2007**, *37*, 151–208.

[42] Salganik, M.J. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health-Bulletin of the New York Academy of Medicine.* **2006**, *83*, I98–I112.

[43] Goel, S.; Salganik, M.J. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences.* **2010**, *107(15)*, 6743–6747, doi.org/10.1073/pnas.1000261107.

[44] Gile, K.J. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association.* **2011**, *106(493)*, 135–146.

[45] Tomas, A.; Gile, K.J. The effect of differential recruitment, non-response and nonrecruitment on estimators for respondent-driven sampling. *Electronic Journal of Statistics.* **2011**, *5*, 899–934, doi.org/10.1214/11-EJS630.

[46] Abdesselam, K.; Verdery, A.; Pelude, L.; Dhami, P.; Momoli, F.; Jolly, A. M. The development of respondent-driven sampling (RDS) inference: A systematic review of the population mean and variance estimates. *Drug and alcohol dependence.* **2020**, *206*, 107702.

[47] Abdul-Quader, A.S. Effectiveness of respondent-driven sampling for recruiting drug users in New York City: findings from a pilot study. *Urban Health Bull.* **2006**, *83(3)*, 459–476, doi.org/10.1007/s11524-006-9052-7.

[48] Wirtz, A.L.; Mehta, S.H.; Latkin, C. Comparison of respondent driven sampling estimators to determine HIV prevalence and population characteristics among men who have sex with men in moscow, russia. *PLoS ONE.* **2016**, *11(6)*, e0155519.

[49] Wejnert, C. An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociological Method.* **2009**, *39(1)*, 73–116, doi.org/10.1111/j.1467-9531.2009.01216.x.

[50] Nesterko, S.; Blitzstein, J. Bias–variance and breadth–depth tradeoffs in respondent-driven sampling. *Journal of Statistical Computation and Simulation.* **2015**, *85(1)*, 89–102, doi.org/10.1080/00949655.2013.804078.

[51] McCreesh, N.; Frost, S.D.W.; Seeley, J.; Katongole, J.; Tarsh, M.N.; Ndunguse, R.; White, R.G. Evaluation of respondent-driven sampling. *Epidemiology.* **2012**, *23(1)*, 138–147, doi.org/10.1097/EDE.0b013e31823ac17c.

[52] Ott, M.Q.; Gile, K.J. Unequal edge inclusion probabilities in link-tracing network sampling with implications for respondent-driven sampling. *Electronic Journal of Statistics.* **2016**, *10(1)*, 1109–1132, doi.org/10.1214/16-EJS1138.

[53] Sperandei, S.; Bastos, L.S.; Ribeiro-Alves, M.; Bastos, F.I. Assessing respondent-driven sampling: A simulation study across different networks. *Social Networks.* **2018**, *52*, 48–55, doi.org/10.1016/j.socnet.2017.05.004.

[54] Barash, V.D.; Cameron, C.J.; Spiller, M.W. Respondent-driven sampling-testing assumptions: sampling with replacement. *Journal of official statistics.* **2016**, *32(1)*, 29–73, doi.org/10.1515/JOS-2016-0002.

[55] Lu, X. Linked ego networks: improving estimate reliability and validity with respondent-driven sampling. *Social Networks.* **2013**, *35(4)*, 669–685, doi.org/10.1016/j.socnet.2013.10.001.

[56] Verdery, A.M.; Merli, M.G.; Moody, J.; Smith, J.A.; Fisher, J.C. Brief report: respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in China. *Epidemiology.* **2015**, *26(5)*, 661–665. doi.org/10.1097/EDE.0000000000000335.

[57] Badowski, G.; Somera, L.P.; Simsiman, B.; Lee, H.R.; Cassel, K.; Yamanaka, A.; Ren, J. The efficacy of respondent-driven sampling for the health assessment of minority populations. *Cancer epidemiology.* **2017**, *50*, 214–220, doi.org/10.1016/j.canep.2017.07.006.

[58] Burlew, K.; Larios, S.; Suarez-Morales, L.; Holmes, B.; Venner, K.; Chavez, R. Increasing ethnic minority participation in substance abuse clinical trials: Lessons learned in the National Institute on Drug Abuse's Clinical Trials Network. *Cultural Diversity and Ethnic Minority Psychology.* **2011**, *17(4)*, 345–356, doi.org/10.1037/a0025668.

[59] Dombrowski, K.; Khan, B.; Moses, J.; Channell, E.; Misshula, E. Assessing respondent driven sampling for network studies in ethnographic contexts. *Advances in Anthropology.* **2013**, *3(1)*, 1–9, doi.org/10.4236/aa.2013.31001.

[60] Firestone, M.; Smylie, J.; Maracle, S.; Spiller, M.; O'Campo, P. Unmasking health determinants and health outcomes for urban First Nations using respondent-driven sampling. *BMJ open.* **2014**, *4(7)*, 1–8, doi:10.1136/bmjopen-2014-004978.

[61] Tyldum, G.; Johnston, L. *Applying respondent driven sampling to migrant populations: Lessons from the field.* Springer, 2014.

[62] Gile, K.J.; Handcock, M.S. Network model-assisted inference from respondent-driven sampling data. *Journal of the Royal Statistical Society.* **2015**, *178(3)*, 619, doi:10.1111/rssa.12091.

[63] Goel, S; Salganik, M.J. Respondent-driven sampling as Markov chain Monte Carlo. *Statistics in medicine.* **2009**, *28*, 2202-2229.

[64] Mendez, M.; Font, J. *Surveying ethnic minorities and immigrant populations.* Amsterdam University Press, 2013. (p. 296).

[65] Feskens, R.; Hox, J.; Lensvelt-Mulders, G.; Schmeets, H. Collecting Data among Ethnic Minorities in an International Perspective. *Field Methods.* **2006**, *18*, 3, 284–304, doi.org/10.1177/1525822X06288756.

[66] Feskens, R.; Hox, J.; Lensvelt-Mulders, G.; Schmeets, H. Nonresponse Among Ethnic Minorities: A Multivariate Analysis. *Journal of Official Statistics.* **2007**, *23*, 3, 387–408.

[67] Erens, B. Designing high-quality surveys of ethnic minority groupsin the United Kingdom. *Surveying Ethnic Minorities and Immigrant Populations: Methodological Challenges and Research Strategies*, Amsterdam University Press: Amsterdam, Holland, 2003.

[68] Kalsbeek, W.D. *Sampling Racial and Ethnic Minorities.* Powerpoint presentation. University North Carolina, 2000. [CrossRef]

[69] Salentin, K. Sampling the Ethnic Minority Population in Germany. The Background to "Migration Background". *Methods, Data, Analyses.* **2014**, *8*, 1, 28, doi.org/10.12758/mda.2014.002.

[70] Sin, C.H. Sampling minority ethnic older people in Britain. *Ageing and Society.* **2004**, *24*, 2, 257-277, doi.org/10.1017/S0144686X03001545.

[71] Shaghaghi, A.; Bhopal, R.S.; Sheikh, A. Approaches to recruiting 'hard-to-reach'populations into research: a review of the literature. *Health promotion perspectives.* **2011**, *1*, 2, 86.

[72] Semaan, S. Time-Space Sampling and Respondent-Driven Sampling with Hard-to-Reach Populations. *Methodological Innovations Online.* **2010**, *5*, 2, 60–75, doi.org/10.4256/mio.2010.0019.

[73] Callegaro, M.; Manfreda, K.L.; Vehovar, V. *Web survey methodology.* London: Sage, 2015.

[74] Baker, R.; Brick, J.M.; Bates, N.A.; Battaglia, M.; Couper, M.P.; Dever, J.A.; Tourangeau, R. Summary report of the AAPOR task force on non-probability sampling. *Journal of survey statistics and methodology.* **2013**, *1*, 2, 90–143.

[75] Van Meter, K. Methodological and design issues: techniques for assessing the representatives of snowball samples. *NIDA Research Monograph.* **1990**, *98*, 51.40, 31–43.

[76] Bhopal, R. Is research into ethnicity and health racist, unsound, or important science?. *Bmj.* **1997**, *314*, 1751, doi.org/10.1136/bmj.314.7096.1751.

[77] Kalsbeek, W.D. Sampling minority groups in health surveys. *Statistics in Medicine.* **2003**, *22*, 9, 1527–1549.

[78] Zambrano, A.K.; Gaviria, A.; Cobos-Navarrete, S.; Gruezo, C.; Rodríguez-Pollit, C.; Armendáriz-Castillo, I.; García-Cárdenas, J.M.; Guerrero, S.; López-Cortés, A.; Leone, P.E.; et al. The three-hybrid genetic composition of an Ecuadorian population using AIMs-InDels compared with autosomes, mitochondrial DNA and Y chromosome data. *Sci. Rep.* **2019**, *9*. doi:10.1038/s41598-019-45723-w.

[79] Moya, J. Migration and the historical formation of Latin America in a global perspective *Sociologias* **2018**, *20*, doi:10.1590/15174522-02004902.

[80] Wade, P. *Race and Ethnicity in Latin America*, 2nd ed.; Pluto Press: London, NY, USA, 2010.

[81] Instituto Nacional de Estadística y Censos. 2010. *El Censo Informa: Educación.* Available online: `http://www.ecuadorencifras.gob.ec/wp-content/descargas/Presentaciones/capitulo_educacion_censo_poblacion_vivienda.pdf` (accessed on 2 September 2020).

[82] Secretaría Nacional de Planificación y Desarrollo. 2017. *Plan Nacional de Desarrollo 2017–2021.* Available online: `http://www.planificacion.gob.ec/wp-content/uploads/downloads/2017/10/PNBV-26-OCT-FINAL_OK.compressed1.pdf` (accessed on 1 August 2020).

[83] Fraze, T.K.; Brewster, A.L.; Lewis, V.A.; Beidler, L.B.; Murray, G.F.; Colla, C.H. Prevalence of Screening for Food Insecurity, Housing Instability, Utility Needs, Transportation Needs, and Interpersonal Violence by US Physician Practices and Hospitals. *JAMA Netw. Open* **2019**, *2*, e1911514, doi:10.1001/jamanetworkopen.2019.11514.

[84] Clapham, D.; Foye, C.; Christian, J. The Concept of Subjective Well-being in Housing Research. *Hous. Theory Soc.* **2018**, *35*, 261–280, doi:10.1080/14036096.2017.1348391.

[85] Ruiz, C.; Hernández-Fernaud, E.; Rolo-González, G.; Hernández, B. Neighborhoods' Evaluation: Influence on Well-Being Variables. *Front. Psychol.* **2019**, *10*, 1736, doi:10.3389/fpsyg.2019.01736.

[86] Chica-Olmo, J.; Sánchez, A.; Sepúlveda-Murillo, F.H. Assessing Colombia's policy of socio-economic stratification: An intra-city study of self-reported quality of life. *Cities* **2020**, *97*, 102560, doi:10.1016/j.cities.2019.102560.

[87] Chica-Olmo, J.; Cano-Guervos, R. Does my house have a premium or discount in relation to my neighbors? A regression-kriging approach. *Socio-Econ. Plan. Sci.* **2020**, 100914, doi:10.1016/j.seps.2020.100914

[88] Boch, S.J.; Taylor, D.M.; Danielson, M.L.; Chisolm, D.J.; Kelleher, K.J. 'Home is where the health is': Housing quality and adult health outcomes in the Survey of Income and Program Participation. *Prev. Med.* **2020**, *132*, 105990, doi:10.1016/j.ypmed.2020.105990.

[89] Ledesma, E.; Ford, C.L. Health Implications of Housing Assignments for Incarcerated Transgender Women. *Am. J. Public Health* **2020**, *110*, 650–654, doi:10.2105/AJPH.2020.305565.

[90] Ramírez-Luzuriaga, M.; Belmont, P.; Waters, W.; Freire, W. Malnutrition inequalities in Ecuador: Differences by wealth, education level and ethnicity. *Public Health Nutr.* **2019**, 1–9, doi:10.1017/S1368980019002751.

[91] Pablo, Q.S.; Paloma, T.L.P.; Francisco, J.T. Energy Poverty in Ecuador. *Sustainability* **2019**, *11*, 6320, doi:10.3390/su11226320.

[92] Vásquez Egas, R. Compañías Incluyentes: ¿una Nueva Receta Para Doblegar a la Pobreza en el Ecuador? Master's Thesis, Universidad Andina Simón Bolívar, Quito, Ecuador, 2009.

[93] Martínez Valle, L. Desarrollo Rural y Pueblos IndíGenas: Las Limitaciones de la Praxis Estatal y de las ONG en el Caso Ecuatoriano. Master's Thesis, FLACSO, Quito, Ecuador, 2002.

[94] Jones, P. Quichua-Castilian bilingualism in the Ecuadorian Sierra. *Early Child Dev. Care* **1994** *102*, 115–138, doi:10.1080/0300443941020109.

[95] Beckett, M.; Firestone, M.A.; McKnight ,C.D. A cross-sectional analysis of the relationship between diabetes and health access barriers in an urban First Nations population in Canada. *BMJ Open* **2018**, *8*, e018272, doi:10.1136/bmjopen-2017-018272.

[96] Firestone, M.; Smylie, J.; Maracle, S.; McKnight, C.; Spiller, M.; O'Campo, P. Mental health and substance use in an urban First Nations population in Hamilton, Ontario. *Can. J. Public Health* **2015**, *106*, e375–e381.

[97] Feskens, R.; Hox, J.; Lensvelt-Mulders, G.; Schmeets, H. Collecting Data among Ethnic Minorities in an International Perspective. *Field Methods* **2006**, *18*, 284–304, doi:10.1177/1525822X06288756.

[98] Chisaguano, S. *La población indígena del Ecuador (Análisis de Estadísticas Socio-Demográficas)*. 2006. Available online: `https://www.acnur.org/fileadmin/Documentos/Publicaciones/2009/7015.pdf` (accessed on 10 August 2020).

[99] Larrea, C.; Torres, F.; López, N.; Rueda, M. *Pueblos Indígenas, Desarrollo Humano y Discriminación en el Ecuador*; Abya Yala: Quito, Ecuador, 2007.

[100] Araki, H. Movimientos étnicos y Multiculturalismo en el Ecuador: Pueblos Indígenas, Afrodescendientes y Montubios. Master's Thesis, University of Kanagawa, Kanagawa, Japan, 2012.

[101] Uquillas, J.; Carrasco, T.; Rees, M. *Exclusión Social y Estrategias de vida de los Indígenas Urbanos en Perú, México y Ecuador*. Available online: `http://repositorio.minedu.gob.pe/handle/123456789/524` (accessed on 10 August 2020).

[102] Guzmán, M.L. Etnicidad y exclusión en Ecuador: una mirada a partir del censo de población de 2001. *Íconos Rev. Cienc. Soc.* **2003**, *17*, 116-132.

[103] Handcock, M.S.; Gile, K.J.; Fellows, I.E.; Neeley, W.W. *Package "RDS"*. **2019** Available online: `https://cran.r-project.org/web/packages/RDS/RDS.pdf` (accessed on 10 August 2020).

[104] Mills, H.L.; Johnson, S.; Hickman, M.; Jones, N.S.; Colijn, C. Errors in reported degrees and respondent driven sampling: implications for bias. *Drug and Alcohol Dependence.* **2014**, *142*, 120–126, doi.org/10.1016/j.drugalcdep.2014.06.015.

[105] Rocha, L.E.; Thorson, A.E.; Lambiotte, R.; Liljeros, F. Respondent-driven sampling bias induced by community structure and response rates in social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* **2016**, *180(1)*, 99–118, doi.org/10.1111/rssa.12180.

[106] Rao, S.; LaRoque, R.; Jentes, E. Comparison of methods for clustered data analysis in a non-ideal situation: results from an evaluation of predictors of yellow fever vaccine refusal in the global TravEpiNet (GTEN) consortium. *Int. J. Stat. Med. Res.* **2014**, *3*, 215-223.

[107] Hubbart, A.E.; Ahern, J.; Fleischer, N.L.; Van der Laan, M.; Lippman, S.A.; Jewell, T.B.; Satariano, W.A. To GEE or not to GEE. *Epidemiology* **2010**, *21*, 467-474.

[108] Sperandei, S.; Bastos, L.S.; Ribeiro-Alves, M.; Reis, A.; Bastos, F.I. Evaluation of Logistic Regression Applied to Respondent-Driven Samples: Simulated and Real Data. *arXiv e-prints.* **2021**, *2101*, 04253.

[109] Bastos, L.S.; Pinho, A.A.; Codeço, C.; Bastos, F.I. Binary regression analysis with network structure of respondent-driven sampling data. *arXiv preprint arXiv.* **2012**, *1206*, 5681.

[110] Besag J. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* **1974**, *36(2)*, 192–225.

[111] Besag, J.; York, J.; Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics.* **1991**, *43*, 1–20.

[112] Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* **2009**, *71*, 319–392.

[113] Manski, C.F. *Partial Identi*

cation of Probability Distributions: Springer Series in Statistics, Springer, New York, United States, 2003.

[114] Whittle, P. *On stationary processes in the plane*, Biometrika 41, 434–449, 1954.

[115] Cressie, N. *Statistics for Spatial Data, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*, New York, United States, 1993.

[116] Dormann, F.C.; McPherson, J.; Araújo, M.; Bivand, R.; Bolliger, J.; Carl, G.; Davies, R.; Hirzel, A.; Jetz, W.; Kissling, W.; et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography.* **2007**, *30(5)*, 609–628.

[117] Yauck, M.; Moodie, E.E.; Apelian, H.; Peet, M.M.; Lambert, G.; Grace, D.; Lachowsky, J.; Hart, T.; Cox, J. Sampling from networks: respondent-driven sampling. *arXiv:2002.05793v2.* **2020**.

[118] Särndal C.E.; Swensson B.; Wretman J. *Model assisted survey sampling.* Springer: Heidelberg, Germany, 1992.

[119] Hansen, M.H.; Hurvitz, W.N. On the theory of sampling from finite populations. *Ann. Math. Stat.* **1943**, *14(4)*, 333-362.

[120] Deville, J.C. *Estimation de la variance pour les enquêtes en deux phases*, Manuscript. Paris, France: INSEE, 1993.

[121] Wolter, K. *Introduction to Variance Estimation*, 2nd Edition, Springer, 2007.