
Una revisión de la metodología de estimación a través de muestreo por cadenas referenciales para las proporciones de una población oculta

A review of the methodology of estimation through respondent-driven sampling for proportions of a hidden population

Edna Carolina Moreno^a
ednamoreno@usantotomas.edu.co

Resumen

Este artículo pone en perspectiva algunos de los resultados más importantes acerca de un nuevo método de muestreo y estimación para poblaciones donde no existe un marco muestral. Tal método es conocido como RDS, por sus siglas en inglés «Respondent-Driven Sampling». Un análisis teórico basado en cadenas de Markov permite mostrar que este método reduce los sesgos generalmente asociados a las muestras por cadenas referenciales, además de producir estimadores asintóticamente insesgados. Se comprobó el potencial del método por medio de una simulación empírica.

Palabras clave: estimador asintóticamente insesgado, población oculta, RDS.

Abstract

This article reviews some of the most important results about a new method of sampling and estimation for populations where there is no sampling frame, this is known as Respondent-Driven Sampling (RDS). A theoretical analysis based on Markov chains shows that this method reduces the bias generally associated with chain-reference samples; in addition the method produces asymptotically unbiased estimators. We demonstrated the potential of the method through an empirical simulation.

Key words: asymptotically unbiased estimator, hidden population, RDS.

^aProfesor. Facultad de Estadística. Universidad Santo Tomás

1. Introducción

En muchas áreas de la investigación social, el problema de obtener una muestra representativa sobre el comportamiento y composición de un determinado grupo social, objeto de estudio, es en la mayoría de los casos resuelto a través de diversas técnicas de muestreo (muestreo aleatorio simple, muestreo aleatorio estratificado, entre otras); sin embargo, en algunas poblaciones con características muy específicas, tales técnicas no son aplicables, ya que ellas requieren que el investigador seleccione miembros para la muestra con una probabilidad conocida de selección, lo cual generalmente exige la existencia o construcción de un *marco muestral*. Para una cantidad importante de grupos o poblaciones sociales, los cuales generalmente constituyen una pequeña porción «oculta» y dispersa de la población, cuyos miembros a menudo suelen ocultar su identidad, como consecuencia de estigmas sociales o problemas con la justicia, *no existe un marco muestral o su construcción es extremadamente costosa y poco práctica* debido a su naturaleza; por ejemplo, grupos sociales, como los drogadictos, los homosexuales, los infectados con VIH, las trabajadoras sexuales, los evasores de impuestos, etc. son difíciles de localizar.

Para estas poblaciones existe la necesidad de una técnica de muestreo que permita hacer estimaciones insesgadas de sus parámetros y obtener muestras representativas. Existen diversas técnicas de muestreo no probabilístico variantes del muestreo por cadenas referenciales, las cuales han generado controversia acerca de su representatividad y validez. En general, una deficiencia de las muestras por cadenas referenciales es que estas pueden ser sesgadas por lo que se llama «**el sesgo de homofilia**», es decir, las referencias que puede hacer un sujeto en particular pueden tener patrones que reflejan tendencias de afiliación con determinados grupos específicos de la población objetivo; por ejemplo, una persona recluta en la muestra personas de sus mismas características, su misma clase social, su mismo nivel de educación o su misma edad, como resultado la muestra final puede llegar a reflejar esta afiliación.

Recientemente, la técnica de muestreo denominada Respondent Driven Sampling (RDS)(Heckathorn 1997) ha demostrado superar las deficiencias de los tipos de muestreo y estimación usualmente empleados, y resulta especialmente eficaz para hacer muestras de poblaciones de este tipo, obteniendo estimaciones asintóticamente insesgadas, y lo que es más importante, teniendo la potencia del muestreo probabilístico.

Se explicará cómo funciona esta técnica de muestreo y estimación en las secciones 2 y 3 y se verificará el potencial del procedimiento en un ejemplo de simulación.

2. Descripción del RDS

El RDS es una técnica de muestreo no probabilístico, basado en la teoría de las cadenas de Markov y la teoría de redes. Produce muestras que son independientes de las características de la muestra inicial de sujetos (también llamada grupo

semilla), además reduce notablemente los sesgos producidos por problemas que suelen presentarse en las poblaciones ocultas como el encubrimiento entre pares, y el sobremuestreo de la población visible, debido a que por la dificultad de localizar sus miembros se tiene tendencia a tomar en la muestra los sujetos más visibles.

Así, el RDS incluye dos componentes: el mecanismo de reclutamiento de sujetos, en el cual la longitud de las cadenas de referencias se produce por una combinación de incentivos, y reclutamiento por cuotas (Heckathorn 1997) y un modelo teórico por medio del cual se calculan los estimadores (Salganik & Heckathorn 2004).

2.1. Recolección de la muestra

Se comienza seleccionando un grupo inicial, «grupo semilla», de miembros de la población objetivo, a través de un contacto preexistente con la población en estudio. Estas semillas son recompensadas, a través de incentivos económicos, por participar en el estudio, y forman lo que se llama «la ola 0 de la muestra». A cada una de estas semillas en la «ola 0» se le provee con una cantidad c de cupones de reclutamiento únicos, y se les pide que entreguen estos cupones a otros miembros de la población objetivo que conozcan. Cuando un nuevo miembro de la población participa en el estudio, se le paga una determinada cantidad a la persona que lo reclutó. De esta manera, a los sujetos se les paga por participar en el estudio y por reclutar a otros miembros en la muestra.

Los reclutados por las personas en la ola 0, constituyen lo que se llama la «ola 1» de la muestra; de la misma manera, a las personas de la ola 1 se les proporciona una cantidad c de cupones de reclutamiento y el proceso continúa de esta manera hasta que se alcance el tamaño de la muestra que se requiere. Hay que destacar que cada cupón es único y puede ser usado posteriormente para determinar los patrones de reclutamiento de la población.

2.2. Descripción teórica de la técnica RDS

Teniendo en cuenta que la muestra obtenida *no es una muestra aleatoria*, se hace necesario un método de estimación especial, específicamente diseñado para obtener estimadores asintóticamente insesgados de la población y que tenga en cuenta las características del proceso de muestreo utilizado.

De esta manera, en lugar de intentar estimar directamente a partir de la muestra los estimadores de la población como lo hacen los métodos de muestreo y estimación usuales, la técnica RDS utiliza la información disponible en la muestra acerca de la *red social*, para obtener estimadores asintóticamente insesgados de las proporciones de la población en los distintos grupos en los que se quiere caracterizar. Estos estimadores se llaman **estimadores de prevalencia** de la población.

Para mostrar de una manera más clara cómo trabaja la técnica RDS se ilustrará para el caso concreto en que la población se clasifica con base en dos características de interés, formando dos grupos, A y B . Lo primero que se hará es es-

tablecer un lenguaje común por medio del cual es posible destacar ciertos supuestos característicos de la red, algunos necesarios para que los estimadores finalmente obtenidos sean asintóticamente insesgados; se mostrará cómo el modelo de reciprocidad permite obtener los dos estimadores de prevalencia de la población; finalmente, se mostrará la razón por la cual son asintóticamente insesgados (Heckathorn 2002).

2.2.1. Estimadores de prevalencia de la población usando la información de la red, a través del modelo de reciprocidad para el caso de dos grupos

El modelo de reciprocidad provee las bases necesarias para poder sacar estimadores válidos de la población a través de muestras RDS. Parte de un rasgo importantísimo en la técnica de muestreo usada y es que en la mayoría de los casos las referencias que hacen los reclutadores son hacia personas cercanas o conocidas, amigos, parejas, hermanos, compañeros de trabajo o de estudio, etc. y se encuentran referencias hacia desconocidos o extraños sólo en una pequeña proporción. Esto indica que en general las relaciones son recíprocas.

Una vez se ha recolectado la muestra, se debe tener un procedimiento que utilice la información contenida en esta para hacer estimaciones sobre la red social y posteriormente hallar los estimadores de prevalencia de la población, esto es, estimar las proporciones de los subgrupos de la población como resultado de clasificarla en base a determinados rasgos distintivos. Con este nuevo objetivo en mente, se establecen una serie de supuestos, algunos determinantes y otros encaminados a facilitar la presentación del tema.

Para lo anterior, se establecerá la siguiente notación:

- T_{AB} : representa el número de lazos de amistad o flechas formados del grupo A al grupo B; un lazo o flecha se interpreta como una amistad o algún tipo de relación social.
- C_{AB} : probabilidad de que al escoger un miembro de un grupo A, se pase al grupo B por medio de un lazo o amistad.
- R_A : número total de lazos o flechas del grupo A.
- d_i : grado de cada persona, reclutado o nodo «i», es decir, el número de miembros de la población que él o ella conoce, o en otras palabras, el número de potenciales reclutados conocidos por la persona.
- D_A : grado promedio de los miembros del grupo
- r_{AB} : número de reclutamientos en la muestra del grupo A en el grupo B.
- n_A : número de elementos de la muestra en el grupo A.
- N_A : número de personas de la población en el grupo A.

Los supuestos requeridos son los siguientes (Salganik & Heckathorn 2004, Sección 6.1):

1. Se asume que todos los lazos de la población son recíprocos. Es decir, que $T_{AB} = T_{BA}$.
2. Se considera sólo el caso del muestreo con reemplazamiento.
3. Se asume que la red de la población oculta forma una componente conectada, es decir, que existe una trayectoria entre cualquier par de personas. Se debe tener cuidado de no aplicar la técnica de estimación RDS en poblaciones en las que no existe un contacto frecuente entre sus miembros y están muy débilmente interrelacionadas, como los evasores de impuestos; es importante destacar que éste es el único supuesto que se requiere de la estructura de la red.
4. Para facilitar la presentación se asume que los reclutadores reciben y usan un solo cupón.
5. Se asume que cuando los reclutados reclutan a otros lo hacen de manera aleatoria entre todas las flechas que rodean al reclutado, es decir, el reclutamiento realizado por una persona se hace de forma aleatoria entre todos los posible lazos de amistad que posee.
6. Finalmente, se supone que las semillas son escogidas con probabilidad proporcional a su grado. Este supuesto será debilitado más adelante, permitiendo una mayor generalidad.

Usando las reglas de la probabilidad condicional, bajo el supuesto de que los nodos en la ola 0 son escogidos con probabilidad proporcional a su grado, entonces la probabilidad de que una amistad o flecha sea escogida en un periodo de reclutamiento es constante e igual para todas las flechas o lazos. Paralelamente, los nodos en todas las olas sucesivas también van a ser escogidos con probabilidad proporcional a su grado (Salganik & Heckathorn 2004, Sección 6.2).

Así, el corazón del modelo de reciprocidad se puede expresar: $T_{AB} = T_{BA}$. Además, obsérvese lo siguiente, R_A se puede expresar como:

$$R_A = \sum_{i \in A} d_i \approx N_A \cdot D_A.$$

Ahora, considérese para una red dada la probabilidad de que al escoger aleatoriamente una amistad comenzando en el grupo A se termine en el grupo B , es posible definir esta probabilidad como:

$$C_{AB} = \frac{T_{AB}}{R_A},$$

Análogamente

$$C_{BA} = \frac{T_{BA}}{R_B}.$$

Así, el número de lazos formados a través de los distintos grupos está dado por:

$$T_{AB} = D_A N_A C_{AB} \quad \text{y} \quad T_{BA} = D_B N_B C_{BA}.$$

Utilizando la hipótesis central de este modelo se obtiene:

$$D_A N_A C_{AB} = D_B N_B C_{BA}.$$

Con objeto de traer a escena las expresiones para las dos proporciones P_A y P_B , se divide a ambos lados de la ecuación por el tamaño total de la población N , y se obtiene:

$$D_A P_A C_{AB} = D_B P_B C_{BA}.$$

En este punto, surge una nueva restricción y es que la suma de las proporciones en las que se subdivide la población debe ser 1. En resumen se obtiene el siguiente sistema de dos ecuaciones:

$$D_A P_A C_{AB} = D_B P_B C_{BA} \tag{1}$$

$$P_A + P_B = 1. \tag{2}$$

Se sustituye $1 - P_A$ por P_B , y despejando P_A , se obtiene una fórmula para la proporción de la población en el grupo A :

$$P_A = \frac{D_B C_{BA}}{D_A C_{AB} + D_B C_{BA}},$$

Análogamente

$$P_B = \frac{D_A C_{AB}}{D_A C_{AB} + D_B C_{BA}}.$$

Estas expresiones para las proporciones de la población basadas en el modelo de reciprocidad requieren únicamente conocer las probabilidades de transición derivadas del análisis de los patrones de reclutamiento, y el reporte del grado personal de cada reclutado (Salganik & Heckathorn 2004, Sección 4).

Evidentemente, para el cálculo de los estimadores de prevalencia se hace necesario estimar el grado promedio de cada grupo, y las probabilidades de transición entre los dos grupos. A continuación se presenta de forma muy resumida la manera de hacerlo y se explica porque cada uno de estos estimadores es asintóticamente insesgado (Salganik & Heckathorn 2004, Sección 6.3), de esta manera la razón entre dos estimadores asintóticamente insesgados también produce un estimador asintóticamente insesgado.

2.2.2. Cálculo de los estimadores necesarios para que los estimadores de prevalencia sean asintóticamente insesgados

Cálculo de los estimadores para C_{AB} y C_{BA}

Debido a que el reclutamiento es una muestra aleatoria de todas las posibles flechas, los estimadores insesgados para C_{AB} y C_{BA} , son:

$$C_{AB} = \frac{r_{AB}}{r_{AA} + r_{AB}}, \quad C_{BA} = \frac{r_{BA}}{r_{BB} + r_{BA}}.$$

Cálculo de los estimadores para D_A y D_B

El método para estimar el grado promedio es a través de un procedimiento estándar, llamado la estimación de Hansen-Hurwitz, que se utiliza cuando el muestreo es con reemplazamiento y las unidades muestrales tienen distinta probabilidad de selección. El procedimiento pesa cada elemento de la muestra por la inversa de su probabilidad de selección, así a las unidades con una pequeña probabilidad de ser seleccionadas se les tiene en cuenta más veces. La ventaja es que este método es conocido por producir estimadores asintóticamente insesgados.

$$D_A^{\hat{h}} = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}}.$$

El numerador y el denominador en $D_A^{\hat{h}}$, son estimadores insesgados por ser obtenidos por el procedimiento de Hansen-Hurwitz, y la razón de dos estimadores insesgados produce un estimador asintóticamente insesgado.

Finalmente, es posible estimar P_A , y P_B :

$$\hat{P}_A = \frac{\hat{D}_B \hat{C}_{BA}}{\hat{D}_A \hat{C}_{AB} + \hat{D}_B \hat{C}_{BA}}, \quad (3)$$

similarmente para \hat{P}_B .

En este punto, es importante observar que es posible demostrar que, sin importar el proceso de selección del grupo semilla, el proceso de muestreo converge

a uno en el cual las personas son escogidas con probabilidad proporcional a su grado, obteniendo la situación que se necesita para que el estimador de C_{AB} sea asintóticamente insesgado (Salganik & Heckathorn 2004, Apéndice C). Esto se logra estableciendo una cadena de Markov en los nodos, esto es: los estados de la cadena corresponden a los distintos nodos, es decir, las N unidades poblacionales y los parámetros corresponden a las distintas olas; por ejemplo, $P[X_3 = 2]$ indica la probabilidad de que en la ola número 3 la cadena se encuentre en el nodo 2. Esta cadena resulta siendo irreducible y ergódica (Heckathorn 2002), lo cual se utiliza para poder demostrar la existencia de las probabilidades de estado estable por medio del siguiente teorema:

Teorema 1. *En una cadena de Markov irreducible ergódica el $\lim_{n \rightarrow \infty} P_{ij}^{(n)}$ existe y es independiente de i . Más aun: $\lim_{n \rightarrow \infty} P_{ij}^{(n)} = \pi_j$, donde las π_j satisfacen de manera única las siguientes ecuaciones de estado estable:*

$$\pi_j = \sum_{i=0}^M \pi_i P_{ij} \quad \text{para } j = 0, 1, 2, \dots, M; \quad \sum_{j=0}^M \pi_j = 1$$

2.2.3. Generalización para el caso de dos grupos o más

Como es de esperarse el modelo 1 también puede extenderse a un sistema con más de dos grupos. Éste requiere resolver un sistema de ecuaciones lineales: la primera de ellas establece que la suma de las proporciones de la población debe ser uno, y las otras reflejan el principio del modelo de reciprocidad para cada par de grupos. Para un sistema de cuatro grupos se tiene:

$$\begin{aligned} 1 &= P_A + P_B + P_C + P_D, \\ P_A D_A C_{AB} &= P_B D_B C_{BA}, \\ P_A D_A C_{AC} &= P_C D_C C_{CA}, \\ P_A D_A C_{AD} &= P_D D_D C_{DA}, \\ P_B D_B C_{BC} &= P_C D_C C_{CB}, \\ P_B D_B C_{BD} &= P_D D_D C_{DB}, \\ P_C D_C C_{CD} &= P_D D_D C_{DC}, \end{aligned} \tag{4}$$

Si los datos se ajustaran perfectamente al modelo de reciprocidad, es decir, si en la red en realidad todos los lazos fueran recíprocos, sería suficiente escoger 4 ecuaciones arbitrariamente y resolver este sub-sistema; sin embargo, como en la realidad este ajuste nunca es perfecto, se encuentran pequeñas diferencias en los resultados de las distintas escogencias. Uno de los métodos más comunes para resolver sistemas de ecuaciones de este tipo es el de los mínimos cuadrados; no obstante, existe un procedimiento alternativo para estimar los parámetros de la población y es el llamado suavizamiento de datos (Heckathorn 2002, p: 23).

El suavizamiento de datos se basa en el principio fundamental del modelo de reciprocidad, esto es, $T_{ij} = T_{ji}$, para todo par de grupos i y j . Este método resuelve el problema de la sobredeterminación del sistema 4; se realiza haciendo un ajuste en la matriz de reclutamiento.

Para este procedimiento es útil ver el proceso de reclutamiento como una cadena de Markov, la diferencia es que en lugar de establecer la cadena sobre los nodos, se hace sobre los distintos grupos de interés de la población, es decir, el proceso de reclutamiento en una muestra por cadenas referenciales «define» una cadena de Markov, en donde el conjunto de parámetros, $t = 0, 1, 2, \dots, n$, de la cadena representan la «ola 0», «ola 1», «ola 2», ..., «ola n » en el proceso de reclutamiento, y el conjunto de estados re-presenta el conjunto de todas las posibles combinaciones de las características de interés para el estudio de la población, esto es, aquellas que subdividen la población objetivo en distintos grupos, cada uno con rasgos o características distintivos. Ahora, una vez concluido el proceso de muestreo por cadenas referenciales, los patrones de reclutamiento de los distintos «grupos» definen dos matrices, una llamada **la matriz de reclutamiento**, que contiene la cantidad absoluta de reclutados por parte de cada grupo en cada uno de los grupos, y otra con las cantidades relativas o proporciones de reclutamientos por cada grupo en cada uno de los otros; y esta matriz es la llamada **matriz de probabilidades de transición**.

Por el momento, se pospondrá el método de suavizamiento de datos para hacer la presentación de las cadenas de Markov en el proceso de reclutamiento, ya que esta parte va a permitir llegar a varias conclusiones importantes y necesarias para poder realizar el procedimiento del suavizamiento de datos y, como ya se dijo, resolver el problema de la sobredeterminación del sistema 4; además, modelar el proceso de reclutamiento como una cadena de Markov permite, como veremos más adelante, proyectar cómo sería la composición de la muestra en ausencia de dos factores de sesgo, el reclutamiento diferencial y las diferencias de los grados promedio. Este resulta siendo el objetivo más importante para realizar este modelamiento y justifica las razones por las cuales utilizar los resultados obtenidos por medio de las cadenas de Markov resulta conveniente para el ajuste en la matriz de reclutamiento.

2.2.4. El proceso de reclutamiento como una cadena de Markov

Es muy razonable pensar en que el proceso de reclutamiento define una cadena de Markov (Heckathorn 1997, p. 182), ya que las características de los reclutadores influyen en las características de los próximos reclutados. Esta relación se puede observar principalmente en lo que se llama el sesgo por homofilia, en el cual el reclutamiento por parte de un persona puede tener patrones que reflejen tendencias de afiliación con determinados grupos específicos de la población. En general, es más probable que una persona con determinada característica social, cultural, etc. sea reclutada por otra dentro de su mismo grupo social o cultural. Teniendo en cuenta que los patrones de reclutamiento sólo dependen de las características de los reclutadores actuales y no de los anteriores, se satisface la propiedad markoviana:

$P[X_{n+1} = j/X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0] = P[X_{n+1} = j/X_n = i]$, lo que significa que la probabilidad de que el siguiente reclutado pertenezca al estado «j», (es decir, tenga la característica j), en la ola $n + 1$, sólo depende del estado en el que se encuentra el reclutador inmediatamente anterior, es decir, en la ola n .

Para aclarar un poco más todo lo mencionado, suponga que se desea estudiar la población de indigentes de la ciudad de Bogotá, y que las características de interés para el estudio son: género (masculino, femenino) y escolaridad (primaria, bachillerato, educación superior); en este caso, el conjunto de estados de la cadena está compuesto por 6 elementos: femenino primaria (estado 1), femenino bachiller (estado 2), femenino superior (estado 3), masculino primaria (estado 4), masculino bachiller (estado 5), masculino superior (estado 6).

Es claro que el vector de probabilidades iniciales \vec{s} definido como: $\vec{s} = (P(X_0 = 0), P(X_0 = 1), \dots, P(X_0 = 6))$, donde $S = \{0, 1, 2, \dots, 6\}$ es el conjunto de estados, representa las proporciones del grupo semilla en los grupos 1, 2, 3, 4, 5, 6. Ahora, la muestra final obtenida proporciona la matriz de probabilidades de transición:

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{16} \\ s_{21} & s_{22} & \cdots & s_{26} \\ \dots & \dots & \dots & \dots \\ s_{61} & s_{62} & \cdots & s_{66} \end{pmatrix}$$

Donde en general el elemento s_{ij} de la matriz representa la proporción de reclutamientos hechos por parte del grupo i . En el grupo j , un simple vistazo le hará ver que es el anteriormente llamado C_{ij} . Supongamos que $s_{41} = 0.70$. Este valor indica que en la muestra el 70% de reclutamientos por parte de grupo de hombres con estudios en primaria fueron hechos dentro del grupo de mujeres con estudios en primaria. Con base en lo anterior, se puede concluir que el muestreo por cadenas referenciales define una cadena de Markov, ya que este proporciona una distribución inicial y una matriz de probabilidades de transición.

El siguiente objetivo es ver que esta cadena de Markov posee exactamente las mismas características que tiene la cadena de Markov establecida para los nodos, es decir, mostrar que es una cadena de markov irreducible y ergódica, y que por lo tanto satisface el teorema 1, página 32. La utilidad de mirar el proceso de reclutamiento como una cadena de Markov radica principalmente en la utilización del teorema 1 que define el vector $\vec{\pi}$ hacia donde converge la matriz de probabilidades de transición, vector que es completamente independiente de la distribución inicial de reclutados. Se denotará este vector de probabilidades de estado estable, para el caso específico, como $\vec{E} = (E_0, E_1, \dots, E_m)$. Donde m , representa el número total de subgrupos de interés de la población, y se llamará el *vector de probabilidades de equilibrio*.

Gracias a que la red forma una componente conectada, todos los estados están completamente ligados, es decir, no existe ningún grupo reclutando únicamente personas dentro de su mismo grupo. Si es así, no es recomendable utilizar la técnica RDS, ya que es importante recordar que una de las características principales de las poblaciones que se tratan con RDS es que tengan fuertes relaciones sociales y

halla una interacción fuerte entre sus miembros. Así, esta característica intrínseca para muchas poblaciones ocultas asegura que la cadena posea una única clase de equivalencia y que la cadena sea irreducible.

Ahora bien, como todos los estados de una cadena de Markov de estado finito son recurrentes, todos los estados de la cadena de Markov definida para los grupos son recurrentes, ya que corresponde a una cadena de Markov de estado finito. Además de esto son aperiódicos, gracias a que es muy probable que en el reclutamiento existan dos olas consecutivas «ola s » y «ola $s + 1$ », tal que el proceso se encuentre en el estado i , en la ola s y en la ola $s + 1$; esto debido a la tendencia de los muestreados a reclutar gente dentro de su misma clase social, cultural o en general dentro de su misma condición. Además, es muy poco probable que un grupo sea completamente heterofílico como para pensar en que no existe ningún par de olas sucesivas en el que la cadena se encuentre en el mismo estado.

De esta manera, se concluye que la cadena de Markov definida para los nodos es una cadena de Markov irreducible y ergódica; y que por lo tanto satisface las hipótesis del teorema 1. Así, si el proceso de reclutamiento, como proceso de Markov, continúa hasta que el equilibrio se alcance, la composición de la muestra final va a ser totalmente independiente del grupo semilla con el cual se inició el proceso de reclutamiento (Heckathorn 2002), y los valores del vector de equilibrio van a corresponder precisamente a las proporciones de reclutados en cada uno de los grupos.

La teoría de las cadenas de Markov proporciona medios para calcular este equilibrio analíticamente:

Suponga que se tienen « m » estados o subgrupos; el vector de equilibrio $\vec{E} = (E_0, E_1, \dots, E_m)$, que representa las proporciones de equilibrio para los grupos $0, 1, \dots, m$, respectivamente, se puede encontrar resolviendo un sistema de m ecuaciones lineales:

$$\begin{aligned} 1 &= E_0 + E_1 + \dots + E_m, \\ E_0 &= s_{00}E_0 + s_{10}E_1 + \dots + s_{m0}E_m \\ E_1 &= s_{01}E_0 + s_{11}E_1 + \dots + s_{m1}E_m, \\ &\dots \\ E_{m-1} &= s_{0m-1}E_0 + s_{1m-1}E_1 + \dots + s_{mm-1}E_m, \end{aligned} \tag{5}$$

Estas proporciones de equilibrio o probabilidades de estado estable contenidas en el vector $\vec{E} = (E_0, E_1, \dots, E_m)$, representan las probabilidades de encontrar el proceso en los estados $0, 1, 2, 3, \dots, m$, después de un número grande de transiciones; por ejemplo, E_2 , representa la probabilidad absoluta de encontrar el proceso de reclutamiento en el estado 2, $p[X_n = 2] = E_2$. Este vector \vec{E} , podría ser un buen candidato para estimar las proporciones de la población en los distintos subgrupos. Sin embargo, los valores obtenidos allí son los esperados «teóricamente», los que se obtienen en ausencia de las dos principales fuentes de sesgo en el proceso de

reclutamiento, y su validez depende del ajuste del proceso de reclutamiento a la cadena de Markov.

Una de las dos grandes fuentes de sesgo en el reclutamiento por cadenas referenciales es la diferencia cuantitativa de los grados promedio de los distintos grupos, porque estos caracterizan de cierta manera la eficiencia del reclutamiento, así si un grupo resulta siendo mucho más eficiente que otro, sus patrones de reclutamiento se van a ver sobrerrepresentados en la muestra. La otra fuente de sesgo es el reclutamiento diferencial en el cual cada grupo tiene tendencias de reclutamiento características, por lo general tienden a reclutar personas dentro de su misma condición social o cultural. Una medida de esto es la homofilia, concepto introducido en Heckathorn (1997) y desarrollado más en detalle en Heckathorn (2002); en general, las diferencias entre la composición de la muestra y el equilibrio son resultado de estas dos fuentes de sesgo. Los siguientes dos teoremas (Heckathorn 2002) tienen gran importancia teórica:

Teorema 2. $\hat{D}_x = \hat{D}_y$ si y solamente si $E = P$; esto es, los estimadores de los grados promedio de los distintos grupos son iguales, si y solo si el vector de equilibrio es igual al vector de proporciones estimadas de la población por la técnica RDS

Teorema 3. Si para cada par de grupos, x y y , $H_x = H_y$, entonces $E = P$.

Para entender el anterior teorema, considérese lo siguiente: La homofilia es una medida que permite cuantificar la tendencia de las personas a reclutar individuos dentro de su mismo grupo, es decir no tiene en cuenta los reclutamientos hechos dentro del grupo resultantes del reclutamiento aleatorio, sino los debidos a esta tendencia; lo anterior produce un sesgo sistemático típico en el muestreo por cadenas referenciales.

Se dice que hay HOMOFILIA PERFECTA cuando los lazos son formados exclusivamente dentro del grupo y se le asigna un valor de «1»; cuando todos los lazos son formados fuera del grupo se le asigna el valor de «-1», similarmente se dice que hay NO HOMOFILIA cuando todos los lazos son hechos sin tener en cuenta el grupo al cual se pertenece, es decir aleatoriamente. En este caso se le asigna un valor de «0». Un concepto paralelo es el de la HETEROFILIA que permite cuantificar la tendencia de las personas a reclutar individuos por fuera de su grupo. La manera de medir el nivel de homofilia en un grupo; se puede ver en Heckathorn (2002).

En general, los grupos con alta homofilia son sobremuestreados, pero este fenómeno se cancela si todos los grupos tienen igual nivel de homofilia; así, el equilibrio converge a la distribución de la población, convirtiéndose en este sentido en un estimador insesgado, que es lo planteado en el teorema 3.

2.3. Suavizamiento de datos

El suavizamiento de datos (Heckathorn 2002, p. 20), ayuda a solucionar el problema de la sobreterminación del sistema 4 y se prefiere este procedimiento y no el de los mínimos cuadrados, porque trae ciertas ventajas adicionales que posteriormente se mencionarán. Se basa en el principio fundamental del modelo de reciprocidad en el que todos los lazos son recíprocos, es decir $T_{ij} = T_{ji}$, para todo par de grupos i, j . Esto implica que el número de reclutamientos hechos por parte del grupo X , que denotaremos RB_X , debería, en teoría, ser igual al número de personas reclutadas de ese mismo grupo, RO_X , esto es, $RO_X = RB_X$, de manera similar se debe tener que para cualquier par de grupos X y Y , el número de reclutamientos a través de cualquier par de grupos deben ser iguales, es decir, $R_{XY} = R_{YX}$.

La idea es proyectar lo que sería la matriz de reclutamiento en ausencia del reclutamiento diferencial, es decir, bajo el supuesto de que todos los grupos reclutan con la misma efectividad, y de manera aleatoria. Esto requiere transformar la matriz de reclutamiento bajo tres condiciones:

- Las probabilidades de transición de la matriz de reclutamiento no deben cambiar, para que los cálculos basados en estas tampoco varíen.
- $RO_X = RB_X$, es decir, el número total de las personas reclutadas dentro del grupo X (la suma de los elementos de la columna X de la matriz de reclutamiento), debe ser igual al número de reclutamientos hechos por parte del grupo X (es decir, la suma de los elementos de la fila X de la matriz de reclutamiento).
- $R_{XY} = R_{YX}$.

El primer paso es ajustar los datos en la matriz de tal manera que el número de sujetos reclutados en un grupo « X », sea igual al número de reclutamientos por parte del grupo « X ». Esto se hace sin afectar los valores de la matriz de probabilidades de transición, de la siguiente manera:

El ajuste para el número de reclutamientos hechos de un grupo i , a un grupo j , R_{ij} es el producto de tres términos S_{ij} , E_i y RB , donde RB es el número total de reclutados que aparecen en la matriz de reclutamiento¹, Para un sistema de M categorías, la matriz de reclutamiento ajustada, R^* , es:

¹El grupo de las semillas no aparece en la matriz de reclutamiento, ya que ellas no tienen reclutador.

$$\mathbb{R}^* = \begin{pmatrix} \hat{S}_{11}\hat{E}_1RB & \hat{S}_{12}\hat{E}_1RB & \cdots & \hat{S}_{1M}\hat{E}_1RB \\ \hat{S}_{21}\hat{E}_2RB & \hat{S}_{22}\hat{E}_2RB & \cdots & \hat{S}_{2M}\hat{E}_2RB \\ \dots & \dots & \dots & \dots \\ \hat{S}_{M1}\hat{E}_M RB & \hat{S}_{M2}\hat{E}_M RB & \cdots & \hat{S}_{MM}\hat{E}_M RB \end{pmatrix}$$

$$= \begin{pmatrix} R_{11}^* & R_{12}^* & \cdots & R_{1M}^* \\ R_{21}^* & R_{22}^* & \cdots & R_{2M}^* \\ \dots & \dots & \dots & \dots \\ R_{M1}^* & R_{M2}^* & \cdots & R_{MM}^* \end{pmatrix}$$

Es claro que la matriz de reclutamiento transformada no altera la proporciones de selección o probabilidades de transición, entre los grupos:

$$\frac{\hat{S}_{ij}\hat{E}_iRB}{\sum_{j=1}^M \hat{S}_{ij}\hat{E}_iRB} = \frac{\hat{S}_{ij}}{\sum_{j=1}^M \hat{S}_{ij}} = \hat{S}_{ij}.$$

Además, la suma de los elementos de la fila i , es igual a la suma de los elementos de la columna i , lo que significa que el número de reclutamientos hechos por el grupo i , RB_i^* , es igual al número de personas reclutadas del grupo i , RO_i^* :

$$RO_i^* = \sum_{k=1}^M \hat{S}_{ki}\hat{E}_kRB = \hat{E}_iRB,$$

Por otro lado:

$$RB_i^* = \sum_{j=1}^M \hat{S}_{ij}\hat{E}_iRB = \hat{E}_iRB;$$

Gracias a que la matriz de probabilidades de transición permanece intacta, los cálculos que se basan en estas proporciones también, es decir el equilibrio. Por otro lado, recordando que los lazos formados deben ser recíprocos para que los datos se ajusten perfectamente al modelo de reciprocidad, entonces esta matriz de reclutamiento debe ser simétrica es decir el número de reclutamientos de un grupo i hacia un grupo j determinado, debe ser igual al número de reclutamientos de j hacia i , y la mejor manera de arreglar este problema es reemplazar cada par de datos ajustados por el promedio entre ellos, esto se hace para todos los pares

posibles obteniendo una matriz de reclutamiento simétrica:

$$\mathbb{R}^{**} = \begin{pmatrix} \hat{S}_{11}\hat{E}_1RB & \frac{(\hat{S}_{12}\hat{E}_1RB)+(\hat{S}_{21}\hat{E}_2RB)}{2} & \dots & \frac{(\hat{S}_{1M}\hat{E}_1RB)+(\hat{S}_{M1}\hat{E}_M RB)}{2} \\ \frac{(\hat{S}_{21}\hat{E}_2RB)+(\hat{S}_{12}\hat{E}_1RB)}{2} & \hat{S}_{22}\hat{E}_2RB & \dots & \frac{(\hat{S}_{2M}\hat{E}_2RB)+(\hat{S}_{M2}\hat{E}_M RB)}{2} \\ \dots & \dots & \dots & \dots \\ \frac{(\hat{S}_{M1}\hat{E}_M RB)+(\hat{S}_{1M}\hat{E}_1RB)}{2} & \frac{(\hat{S}_{M2}\hat{E}_M RB)+(\hat{S}_{2M}\hat{E}_2R)}{2} & \dots & \hat{S}_{MM}\hat{E}_M RB \end{pmatrix},$$

De esta manera, la nueva matriz de reclutamiento es completamente compatible con el modelo de reciprocidad, dejando la libertad de resolver el sistema 4 , seleccionando 4 ecuaciones arbitrariamente. Por ejemplo:

$$\begin{aligned} 1 &= \hat{P}_1^{**} + \hat{P}_2^{**} + \hat{P}_3^{**} + \dots + \hat{P}_M^{**}, \\ \hat{P}_1^{**} \hat{D}_1 \hat{S}_{12}^{**} &= \hat{P}_2^{**} \hat{D}_2 \hat{S}_{21}^{**}, \\ \hat{P}_1^{**} \hat{D}_1 \hat{S}_{13}^{**} &= \hat{P}_3^{**} \hat{D}_3 \hat{S}_{31}^{**}, \\ &\vdots \\ \hat{P}_1^{**} \hat{D}_1 \hat{S}_{1M}^{**} &= \hat{P}_M^{**} \hat{D}_M \hat{S}_{M1}^{**}. \end{aligned} \quad (6)$$

Este sistema de ecuaciones proporciona los medios para calcular los estimadores de la población para un sistema con M categorías, resolviendo un sistema de M ecuaciones con M incógnitas. Además, después de que se tiene la matriz de reclutamiento suavizada, esta se convierte en la base para realizar todos los cálculos, así todos los términos dependientes de la matriz cambian, entre esos el equilibrio. El suavizamiento de datos se utilizará para el caso de tres grupos o más; en el caso de solo dos grupos, el sistema es sencillo de resolver.

La ventaja del suavizamiento de datos es que además de resolver el problema de la sobredeterminación del sistema, este procedimiento reduce el número de términos a partir de los cuales se calculan los estimadores de la población a la mitad; así, el estimador calculado se vuelve más eficiente. Otra ventaja es que la matriz transformada preserva un rasgo importantísimo y es que si los grados son iguales, el estimador de la población es igual al de equilibrio.

3. Simulación de una población

En estas dos últimas secciones se hará uso de un programa que se creó para simular una población y una muestra con las características del muestreo por cadenas referenciales. Este programa permite simular una población de tamaño arbitrario, a cuyos miembros les asigna dos o tres características ($\{1,2\}$ o $\{1,2,3\}$) según se

necesite, es decir, a cada miembro « i », $i = \{1, 2, 3, \dots, N\}$, donde N es el tamaño de la población, se le clasifica en un grupo « j », $j = \{1, 2, 3\}$. De igual manera, a cada miembro de la población se le asigna un grado, que depende del grupo del cual es miembro. Los grados de cada grupo se distribuyen de manera exponencial con una media que se elige de manera arbitraria; con base en este grado, se les asignan sus respectivas amistades de manera completamente aleatoria, esto es, si al miembro de la población « i » se le asignó un grado « $d_i = 20$ », a este miembro se le asignan 20 individuos de la población aleatoriamente para conformar la red. Para la muestra se inicia con un grupo semilla arbitrario, en todas las simulaciones se tomó un número de cupones igual a 3; así, un individuo puede reclutar 0, 1, 2 o 3 personas, que son elegidas aleatoriamente entre las amistades asignadas.

Es necesario recalcar que las poblaciones simuladas con este programa tienden a tener una alta componente de aleatoriedad por la forma en que fue construido el programa, así las proporciones en la muestra resultarían siendo un muy buen estimador de los parámetros de la población; sin embargo, una de las características reales del muestreo por cadenas referenciales es que los grupos con grados altos resultan siendo sobremuestreados, es decir, en la práctica la muestra no es aleatoria, por lo que las proporciones de la muestra no son buenos estimadores, y gracias a este hecho fue que se generó la técnica de estimación RDS. De cualquier manera, la simulación sirve para verificar de cierta manera el nivel de eficiencia de la técnica, teniendo cuidado de que la técnica se planteó pensando en una población con las características ya mencionadas en las anteriores secciones, así no resultará muy útil al aplicarla, primero, a una muestra aleatoria, y segundo, a una población donde no se tuvo en cuenta el grado de cada persona al asignar las respectivas amistades, en el sentido de que una persona, entre mayor sea su grado, más probabilidades tiene de ser seleccionado dentro del muestreo, reflejándose así tendencias de homofilia o heterofilia.

Para ilustrar la teoría expuesta se simuló en principio una población y una muestra de 10000 y 1215 miembros respectivamente, a cada uno de los miembros de la población se les asignó uno de tres rasgos o características posibles entre «1,2,3», que pueden representar en un caso particular estrato 1, estrato 2, estrato 3, con las proporciones mostradas en la tabla 1; igualmente, se les asignó un grado que depende del grupo al cual se es miembro: por ejemplo, a los miembros del grupo «1» se les asignaron grados que se distribuyen exponencialmente con parámetro $\mu_k = 10$. La red social se logró asignando de manera aleatoria a cada uno de los miembros de la población un número de amistades igual al grado asignado. A partir de esta red se tomó la muestra iniciando con un grupo semilla de 5 personas, y con 3 cupones para cada reclutador, teniendo en cuenta que una persona puede o no reclutar, es decir, puede no utilizar ninguno de los cupones, solo 1, o todos; los siguientes son los datos completos para la población y la muestra simulada, incluso se muestran aquellos que son necesarios para el cálculo del grado promedio de cada grupo.

Tabla 1: *Características de la población y la muestra simuladas.*

DATOS	GRUPO 1	GRUPO 2	GRUPO 3
N_k	6368	3238	394
PN_k	0,637	0,324	0,039
n_k	807	356	52
Pn_k	0,664	0,293	0,043
μ_k	10	20	30
$\sum_{i=1}^{n_k} d_i$	8700	7052	1549
$\sum_{i=1}^{n_k} \frac{1}{d_i}$	218,468	62,044	6,079

Donde:

N_k := «Número absoluto de miembros de la población que pertenecen a cada uno de los grupos».

PN_k := «Proporción de miembros de la población en cada uno de los grupos».

n_k := «Número absoluto de miembros de la muestra que pertenecen a cada uno de los grupos».

Pn_k := «Proporción de miembros de la muestra en cada uno de los grupos».

μ_k := «Parámetro de la distribución exponencial asignado a cada grupos en la simulación».

$\sum_{i=1}^{n_k} d_i$:= «Sumatoria de los grados del grupo k , $k = 1, 2, 3$ ».

Es importante aclarar que esta muestra se simuló sin tener en cuenta los supuestos mencionados para el modelo, 2.2.1.

El propósito de esta población simulada es simplemente ejemplificar algunos de los cálculos utilizados en el procedimiento de estimación de las proporciones de la población, también permite hacerse una idea de la efectividad del proceso de estimación. Los resultados se obtuvieron utilizando el programa «RDSAT», disponible en la página de internet www.respondentdrivensampling.org.

3.1. Cálculo del estimador del grado promedio de cada uno de los grupos

Si se recuerda bien, para el proceso de estimación de las proporciones de la población es absolutamente indispensable el cálculo de un estimador asintóticamente insesgado para el grado promedio de cada uno de los grupos. Este estaba dado por:

$$\hat{D}_A = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}} \quad (7)$$

Así, con base en los datos suministrados por la tabla 1, los estimadores de los grados para los distintos grupos se presentan a continuación junto con el promedio de los grados en cada grupo, recordando, claro está, que el simple promedio no es un buen estimador:

Tabla 2: *Estimadores de los grados promedio.*

GRUPO	PROMEDIO ARITMETICO DE LOS GRADOS	ESTIMADOR UTILIZANDO LA FORMULA 7
GRUPO 1	10,765	3,693
GRUPO 2	19,796	5,709
GRUPO 3	29,788	8,552

La fórmula para estimar el promedio de los grupos, controla de cierta manera el efecto que tiene sobre los estimadores el hecho de que algunos reclutados tengan grados muy altos. El estimador para el grado presentado aquí es un caso particular del estimador del grado ajustado presentado en Heckathorn (2007); sin embargo, conserva la idea esencial de que cada uno de los reclutados sean pesados por el recíproco de su grado, esto debido a que como ya se ha mencionado, un reclutado con un grado alto tiene una alta probabilidad de ser incluido en la muestra ya que hay más trayectorias que conducen a él. Se hace esto para controlar esta fuente de sesgo.

También se debe aclarar que las diferencias de los estimadores de las proporciones de la población utilizando el estimador del grado presentado en el presente documento y utilizando el estimador propuesto en el artículo del 2007 por Heckathorn, no son muy significativas.

3.2. Suavizamiento de datos y resumen de los resultados obtenidos

Toda la información contenida en esta sección se obtuvo al correr el análisis de los datos de la muestra con el programa «RDSAT», que genera el análisis utilizando la técnica RDS, además da las proporciones de la muestra y estimaciones de la homofilia en la población, entre otros. Se podrán comparar los estimadores de las proporciones de la población obtenidos por el método de suavizamiento de datos y el de los mínimos cuadrados (como métodos para resolver el sistema 4), con los parámetros verdaderos de la población. Los siguientes son los resultados:

Tabla 3: *Cantidades absolutas y proporciones de reclutamientos.*

GRUPO	Datos	1	2	3
1	RB_{ij}	302	135	21
	S_{ij}	0,659	0,295	0,046
2	RB_{ij}	153	64	10
	S_{ij}	0,674	0,282	0,044
3	RB_{ij}	17	9	0
	S_{ij}	0,6538	0,34615	0

Donde:

$RB_{ij} :=$ «Cantidad de reclutamientos del grupo i al grupo j .»

$S_{ij} :=$ «Proporción de reclutados del grupo i , al grupo j .»

Los datos de la matriz de reclutamiento ajustada R^* :

Tabla 4: *Matriz de reclutamiento ajustada R^* .*

GRUPO	1	2	3
1	311.035	139.038	21.628
2	140.520	580779	9184
3	20.146	10.665	0.0

Al volverla simétrica obtenemos la matriz R_{**} :

Tabla 5: *Matriz de reclutamiento ajustada simétrica R_{**} .*

GRUPO	1	2	3
1	311.035	139.779	20.887
2	139.779	58.779	9.925
3	20.887	9.925	0.0

Tabla 6: *Resumen de resultados.*

DATOS	GRUPO 1	GRUPO 2	GRUPO 3
P. población	0,637	0,324	0,039
P.muestra	0,664	0,293	0,043
P. mínimos cuadrados	0,762	0,216	0,022
P. Suavizamiento	0,761	0,218	0,021
Grupo semilla	2	3	0
Homofilia	-0,135	0,084	-1,0
Intervalos de confianza para los estimadores	0,727-0,797	0,181-0,25	0,013-0,032

En primer lugar, los estimadores de las proporciones de la población hipotética en los grupos «0,1,2,3», comparados con los parámetros verdaderos de la población reflejan la importancia de estudiar el ajuste de los datos a los 6 supuestos, y de analizar las posibles fuentes de sesgo. En este caso se tomó un tamaño de muestra de 1215 personas, bastante grande para un población de 10.000 habitantes y como consecuencia las proporciones de la muestra resultan siendo un mejor estimador. Sin embargo, en la práctica el tamaño de las muestras es mucho menor en proporción que el tamaño de la población; aún así, los estimadores por suavizamiento de datos aunque no muy satisfactorios, reflejan el potencial del procedimiento. Además de esto, es posible ver que los estimadores por mínimos cuadrados se diferencian como máximo en 0,002 décimas de los estimadores por suavizamiento de datos 6.

Una posible fuente de sesgo en los estimadores obtenidos en este ejemplo puede haber sido el hecho de que el grupo 3 es completamente heterofílico, hecho que no se presenta en la realidad.

Otro resultado importante obtenido en el programa RDSAT son los intervalos de confianza de los estimadores para cada uno de los grupos, estos intervalos se obtienen mediante el procedimiento de Bootstrap, el cual usa la muestra para generar un conjunto de muestras replica del mismo tamaño de la muestra original, de las cuales se obtiene un conjunto de estimadores replica, y partir de ahí es posible construir intervalos de confianza alrededor del estimador original.

3.3. Un breve análisis de sensibilidad

Para el siguiente análisis se simularon tres poblaciones, en las cuales se clasificaron a sus miembros en base a dos características de interés 1 y 2, con las siguientes características comunes:

Tabla 7: *Características de las poblaciones simuladas para el análisis de sensibilidad.*

Tamaños de las poblaciones	10000
P. población en el grupo 1	0.6963
P. población en el grupo 2	0.3037
Tamaño de muestra	500
Semillas	Nodos 1,2,3,4,5
Número de cupones	3

La idea del análisis de sensibilidad es estudiar cómo varían los estimadores de la población para los grupos «1» y «2», en poblaciones que comparten las características dadas en la tabla 7, es decir, las tres poblaciones tienen 10000 habitantes exactamente, con la misma proporción de miembros en los grupos «1» y «2». Se toma en cada una de ellas un grupo semilla conformado por los nodos «0,1,2,3,4,5», y se conforma una muestra; la característica distintiva de cada población son los grados de la distribución exponencial² presentados en la siguiente tabla:

Tabla 8: *Parámetros de la distribución exponencial utilizada para simular los grados.*

POBLACIÓN	GRADO EXPONENCIAL GRUPO 1	GRADO EXPONENCIAL GRUPO 2
POBLACION 1	5	20
POBLACION 2	10	20
POBLACION 3	10	10

²Hay evidencia, en estudios ya realizados de que las distribuciones de probabilidad de los grados en poblaciones ocultas siguen una distribución exponencial.

Se corrieron las tres muestras generadas y se obtuvieron los resultados que muestra la tabla 1; para efectos de comparación se anexan las proporciones verdaderas y la media de la distribución exponencial de la población:

Tabla 9: *Resumen de resultados de las simulaciones.*

DATOS	.	POB 1	POB 2	POB 3
Proporción poblacional	«1»	0.6963	0.6963	0.6963
	«2»	0.3037	0.3037	0.3037
P.muestra	«1»	0.668	0.687	0.723
	«2»	0.332	0.313	0.277
P. mínimos cuadrados	«1»	0.8225	0.775	0.742
	«2»	0.175	0.225	0.258
P. Suavizamiento	«1»	0.825	0.775	0.742
	«2»	0.175	0.225	0.258
Media de la distribución exponencial	«1»	5	10	10
	«2»	20	20	10

Es posible observar que se obtienen mejores estimadores por el método de suavizamiento de datos o por el método de los mínimos cuadrados, cuando no existe el sesgo resultado de las diferencias entre los grados promedio de los distintos grupos.

Para la primera población en la cual el grado de la distribución exponencial en el grupo 1 es de 5 y en el grupo 2 es de 20, se observa que se obtienen muy malos estimadores comparados con las proporciones verdaderas de la población. Los estimadores de la población por el método de suavizamiento de datos son del 82.5%, y 17.5% para el grupo «1» y «2» respectivamente comparado con las proporciones verdaderas 69.03% y 30.37%, mientras que para la segunda población los estimadores mejoran un poco debido posiblemente a que la diferencia entre los grados promedio de los dos grupos no es muy grande. En la tercera población se obtienen resultados mucho más satisfactorios, con estimadores de 74.2% y 25.8%. Además, según el teorema 2, página 36, podríamos concluir que estas proporciones son iguales a las del equilibrio.

Lo que se pretende resaltar aquí es la importancia de encontrar técnicas que controlen y midan las principales fuentes de sesgo en esta técnica de estimación, las diferencias en los grados promedio y el reclutamiento diferencial. La conclusión principal de esta sección es que entre más difieran los grados promedio de los distintos grupos en la población los estimadores tendrán un mayor sesgo.

4. Conclusiones

En muchas áreas de la investigación se tiene la necesidad de estimar proporciones de poblaciones dispersas, en las que sus miembros por lo general no son reconocidos abiertamente, pero que poseen una densa red social de amistades; para estas poblaciones se han venido utilizando muestras por cadenas referenciales que no

gozan de mucha credibilidad por parte de los científicos por considerarse muestras por conveniencia, además porque, al no ser muestras aleatorias, no es posible hacer inferencias sobre ellas. Sin embargo, uno de los objetivos del presente documento es rescatar el potencial de estas muestras por cadenas referenciales e ilustrar una aproximación alternativa del muestreo de poblaciones ocultas, propuesta por Douglas D. Heckathorn. Como se vio, dicha aproximación goza de la propiedad de que bajo ciertos supuestos ya mencionados, sus estimadores resultan siendo asintóticamente insesgados.

Por otro lado, se demostró con la ayuda de la teoría de las cadenas de Markov, que si el proceso de reclutamiento como proceso de Markov continua hasta que el equilibrio se alcance, la composición de la muestra final es totalmente independiente del grupo semilla; de esta manera, se soluciona uno de los grandes conflictos del muestreo por cadenas referenciales.

Como bien puede observarse a lo largo del presente documento la técnica RDS, proporciona información no sólo de las proporciones de la población, sino también de su componente social, de la manera como se relacionan sus miembros o lo que es lo mismo, de su red social de amistades con características como la homofilia. Esta información puede ser extremadamente útil en problemas de salud pública como en el estudio de dispersión de enfermedades.

Además de esto, esta técnica se caracteriza por ser mucho más económica, rápida y efectiva que otras técnicas de muestreo utilizadas, lo cual resulta siendo una ventaja muy significativa.

Lo último que queda por resaltar son algunos resultados que no se mencionan en el presente documento, y que merecen especial atención por parte del lector interesado. Dichos resultados giran en torno a aspectos como al tamaño de la muestra adecuado, el cálculo de intervalos de confianza de los estimadores y las técnicas para controlar y medir las principales fuentes de sesgo de esta técnica de estimación.

Naturalmente, es conveniente verificar cada uno de los resultados obtenidos por simulación, y muy importante realizar análisis de sensibilidad variando tamaños de muestra, niveles de homofilia, heterofilia, número de cupones utilizado; para descubrir patrones que luego puedan ser verificados analíticamente. De igual manera, es muy importante el cálculo del número de olas necesarias para alcanzar el equilibrio, ya que esta es una de las bases fundamentales en el procedimiento de estimación.

Merece también especial atención estudiar los medios para el análisis de variables continuas, y el control del sesgo debido al reclutamiento diferencial (Heckathorn, 2007).

Agradecimientos

Agradezco a Leonardo Sánchez por la asistencia computacional.

Recibido : 24denoviembrede2009

Aceptado : 10demarzode2010 (8)

Referencias

- Heckathorn, D. D. (1997), 'Respondent-driven sampling: A new approach to the study of hidden populations', *Social Problems* **44**(174-199).
- Heckathorn, D. D. (2002), 'Respondent-driven sampling ii: Deriving valid population estimates from chain-referral samples of hidden populations', *Social Problems* **49**(11-34).
- Heckathorn, D. D. (2007), 'Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment', *Sociological Methodology* **37**(151-208).
- Salganik, M. J. & Heckathorn, D. D. (2004), 'Sampling and estimation in hidden populations using respondent-driven sampling', *Sociological Methodology* **34**(193-239).