ESTIMACIÓN DE LA PROPORCIÓN DE UNA SUBPOBLACIÓN OCULTA A TRAVÉS DE MUESTREO POR "BOLA DE NIEVE" ESTRATIFICADO

Por

William Benedicto Sarmiento Rondón

Tesis sometida en cumplimiento parcial de los requerimientos para el grado de

MAESTRÍA EN CIENCIAS

en

ESTADÍSTICA

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ

Diciembre, 2008

Aprobada por:

Julio C. Quintana Díaz, Ph.D Presidente, Comité Graduado	Fecha
Dámaris Santana Morant, Ph.D Miembro, Comité Graduado	Fecha
Edgardo Lorenzo González, Ph.D Miembro, Comité Graduado	Fecha
Walter Díaz Rodríguez, Ph.D Representante de Estudios Graduados	Fecha
Julio C. Quintana Díaz, Ph.D Director del Departmento	Fecha

Abstract of Thesis Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Science

ESTIMATION OF THE PROPORTION OF A HIDDEN POPULATION BY STRATIFIED "SNOWBALL" SAMPLING

By

William Benedicto Sarmiento Rondón

December 2008

Chair: Julio C. Quintana Díaz
Major Department: Mathematical Sciences

The inference process is inefficient when employing conventional sampling techniques to sample from hidden populations and those of difficult access. The method of sampling by reference or "Link-Tracing" ("Snowball Sampling") permit provides additional information by using the existing relations or ties among its members, starting from an initial selected sample. In this context the population is presented as a graph, whether directed o not, where each individual is a node that describes its characteristic and each edge, the existing relation between each one of them. From a sociological perspective, human beings form social groups based on particular criterion that could potencially affects parameter estimation. From this point of view, we will apply "Link-Tracing", by selecting an initial stratified random sampling, instead of an initial simple random sample, in order to increase the precision of our estimates, because the homogeneity intra strata. The objective is to estimate the size and proportion of a hidden population, from an initial sample S_0 and S_1, identifying each individuals characteristic, and considering the existing relationships, within each stratum. Our second objetive was to develop and implement algorithms

in the statistics software **R**, to be applied to real and simulated data and compare these results to those obtained from conventional sampling procedures.

Key Words: Link-Tracing, Hidden Population, Graph Sampling, Stratified Sampling, Snowball Sampling, Adaptive Sampling.

Resumen de Tesis presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
requerimientos para el grado de Maestría en Ciencias

ESTIMACIÓN DE LA PROPORCIÓN DE UNA SUBPOBLACIÓN OCULTA A TRAVÉS DE MUESTREO POR "BOLA DE NIEVE" ESTRATIFICADO

Por

William Benedicto Sarmiento Rondón

Diciembre 2008

Consejero: Julio C. Quintana Díaz
Departamento: Ciencias Matemáticas

En poblaciones ocultas y de difícil acceso el proceso de inferencia es ineficiente cuando se emplean técnicas de muestreo convencionales. El muestro por referencia o "Rastreo por vínculos" ("Bola de Nieve") permite obtener más información, usando las relaciones existentes entre sus miembros a partir de una muestra inicial. En este contexto la población se representa como un grafo, donde cada individuo es un nodo que describe su característica y cada arco la relación existente entre ellos. Desde la perspectiva sociológica, el ser humano tiende a formar estratos bajo algún criterio de asociación, que pueden influir en la inferencia de un parámetro en particular. Desde este enfoque, aplicar la técnica de "Rastreo por vínculos" escogiendo la muestra inicial de la población completa, en lugar de seleccionar una muestra inicial considerando los estratos, no sería apropiado si la característica bajo estudio define una división natural en la población, con una homogeneidad interna establecida. El objetivo es estimar la proporción y el tamaño de una subpoblación oculta, a partir de una muestra inicial S_0 y de los contactos S_1, identificando la característica de cada individuo y considerando las relaciones existentes, en cada estrato. Nuestro

segundo objetivo es desarrollar e implementar en el programa estadístico **R** algoritmos, aplicado a datos reales y simulados, para comparar los resultados obtenidos con los estimadores de muestreo convencional.

Palabras claves: Rastreo por vínculos, Población Oculta, Muestreo en Grafos, Muestreo Estratificado, Bola de Nieve, Muestreo Adaptativo.

A Luz Verónica, retoño de vida que me ha brindado mi DIOS.

A mi madre bella y hermosa con mucha alegría y cariño, que dió todo lo que tenía por mí.

AGRADECIMIENTOS

En primer lugar quiero dar las gracias a Dios, a su Hijo Jesuscristo y al Espíritu Santo por ser esa fuerza invisible y real que motiva con más intensidad en los momentos difíciles.

A Luz Marina, mi pedazo de cielo que se convirtió en un inmenso cielo.

A mis hermanos Tere, Martina y Jaime que siempre están presentes en cualquier momento.

Al doctor Julio César Quintana Díaz que siempre confió y creyó en mí desde el primer día de clase. Gracias por todos su consejos y observaciones en la realización de este trabajo.

A los doctores Dámaris Santana, Edgardo Lorenzo y Walter Díaz miembros de mi comité graduado y representante de Estudios Graduados.

A Julián García por todas sus palabras, actitudes y cualidades de lo que significa ser un amigo.

A los doctores Steve Muth y John Potterat por facilitarme y permitir usar los datos del estudio Colorado Springs.

Y por último mis agradecimientos a David y todas aquellas personas que permitieron que se llegara al fin del camino.

TABLA DE CONTENIDO

página

RESUMEN EN INGLÉS ii

RESUMEN EN ESPAÑOL iv

AGRADECIMIENTOS viii

LISTA DE TABLAS xi

LISTA DE FIGURAS xii

LISTA DE ABREVIATURAS xiii

LISTA DE SÍMBOLOS xiv

1 INTRODUCCIÓN 1

2 PRELIMINARES 3

2.1 Poblaciones ocultas o escasas 3
2.2 Tipos de diseño de muestreo adaptativos 5
2.2.1 El diseño “Rastreo por vínculos” 7
2.2.2 Otros tipos de muestreo adaptativo 10
2.2.3 La teoría de grafos en “Rastreo por vínculos” 11

3 TRABAJOS PREVIOS 13

4 INFERENCIA DEL DISEÑO ESTRATIFICADO CON RASTREO DE VÍNCULOS 21

4.1 Modelo del grafo y diseño del muestreo 22
4.2 Formulación teórica de Frank, Thompson y Chow 24
4.2.1 Relación de los vínculos con el valor de los nodos. 24
4.2.2 Inferencia del diseño Rastreo de vínculos 25
4.3 Una demostración numérica con datos reales 32

5 RESULTADOS DEL MÉTODO ESTRATIFICADO-BOLA DE NIEVE . 35

5.1 Simulaciones y resultados 38
5.2 Resultados obtenidos DATA 1. 39
5.3 Resultados obtenidos DATA 2. 45

6 CONCLUSIONES Y TRABAJOS FUTUROS 52

6.1 Conclusiones 52
6.2 Contribuciones 53
6.3 Trabajos futuros 53

REFERENCIAS BIBLIOGRÁFICAS 54

APÉNDICES 58

A ALGORITMOS EN R 59

LISTA DE TABLAS

Tabla página

4–1 Estimación en Colorado Springs 33

5–1 Proporción 0.1, grafo de 500 nodos 39

5–2 Proporción 0.05, grafo de 500 nodos 40

5–3 Proporción 0.02, grafo de 500 nodos 40

5–4 Error cuadrático medio de los estimadores en DATA1 42

5–5 Proporción 0.1, grafo de 1000 nodos 45

5–6 Proporción 0.05, grafo de 1000 nodos 46

5–7 Proporción 0.025, grafo de 1000 nodos 46

5–8 Proporción 0.01, grafo de 1000 nodos 47

5–9 Error cuadrático medio en los estimadores de DATA2 48

LISTA DE FIGURAS

Figura página

5–1 Grafo de DATA1 con 500 nodos en modo random 41
5–2 Grafo de DATA1 con 500 nodos en modo estándar 43
5–3 Muestra S_0=50 DATA1 43
5–4 Muestra $S_0 \times S_0$ Proporción=0.1 44
5–5 Muestra S_0=25 Proporción=0.1 44
5–6 Grafo de DATA2 con 1000 nodos en modo random 47
5–7 Muestra S_0=50 DATA2 49
5–8 Muestra $S_0 \times S_0$ Proportion=0.1 49
5–9 Muestra S_0=50 Proportion=0.1 50

LISTA DE ABREVIATURAS

θ_1	Estimador Bayes individuos con la característica.
θ_{1pr}	Estimador Bayes Predictor individuos con la característica.
prop	Estimador usual de Proporción.
mse	Error cuadrático medio.

LISTA DE SÍMBOLOS

$\hat{\theta}_1$	Estimador de la proporción de la subpoblación con la característica de interés.
$\hat{\beta}_2$	Estimador de la probabilidad de un vínculo mutuo entre dos individuos con la característica.

CAPÍTULO 1
INTRODUCCIÓN

La estadística es un ciencia que ha generado y ha facilitado el análisis y manejo de datos obtenidos a través de experimentos, tratamientos y demás situaciones del contexto científico.

La recolección y análisis de datos es de gran importancia para el estudio de alguna situación en particular. La forma y técnica utilizada para la obtención de ellos repercute en los resultados que se obtienen, afectando su veracidad, confiabilidad y consistencia en la decisión a tomar.

Dependiendo del estudio a realizar, de su naturaleza, de la clase de la población y demás aspectos, la recolección de los datos se puede convertir en un desafío para el investigador a causa de diferentes factores.

Cuando se tiene el objetivo claro de lo que se quiere investigar, y se han definido e identificado el marco muestral, la metodología, la técnica de muestreo, el tamaño de muestra, el instrumento de recopilación de datos y demás elementos; el estudio se hace fácil y no genera reto alguno para la teoría y el análisis estadístico que el tema requiere. En contraste, cuando el marco muestral no existe y por consiguiente el tamaño de la población no se conoce, teniendo en cuenta que el conocimiento previo de la población es mínimo [1], son razones que dificultan e impiden al investigador llevar a cabo su próposito específico en particular con poblaciones de características ilegales, sensitivas o estigmatizadas.

En la sociedad se presentan conductas que de una u otra forma crean grupos que no quieren ser señalados, pues sus comportamientos no son aceptados socialmente,

así tengan o no el respaldo de las leyes y, por lo tanto prefieren mantenerse ocultos y asociados entre sí [2]. En estas situaciones la teoría de muestreo convencional no genera resultados óptimos y consistentes para la estimación de diversos parámetros de tales poblaciones [3], por lo cual se hace necesario aplicar teorías y métodos estadísticos que produzcan los estimados confiables que se desea y que permitan hacer inferencia sobre las mismas. Teniendo en cuenta que existen relaciones entre los miembros de estas poblaciones, se emplea un diseño de muestreo adaptativo denominado "Rastreo por vínculos", que consiste en obtener información y nuevos contactos a partir de una muestra inicial seleccionada, el cual en estos casos es la forma más práctica y viable de obtener una muestra grande y representativa [4].

El esquema general que se siguió en este trabajo de investigación fue utilizar conceptos y términos de la teoría de grafos para la estimación de parámetros de poblaciones ocultas o escasas [5], examinar las distintas técnicas de muestreo en este tipo de poblaciones, y desarrollar diferentes estimadores. En trabajos previos se ha utilizado la muestra inicial aleatoria de la poblacion completa para obtener estos estimadores. En esta tesis se incorporó la aplicación de una muestra inicial estratificada con el propósito de utilizar la homogeneidad existente entre estratos para obtener mejor precisión de los estimadores, es decir se combinó el muestreo estratificado con "Rastreo por vínculos". A la vez se desarrollaron los algoritmos en **R** para las diversas variaciones consideradas en el muestreo estratificado con relación a los tipos de arcos existentes. Por último mostrar y comparar la eficiencia de los resultados obtenidos, en este tipo de muestreo propuesto con los métodos convencionales de muestreo, usando datos reales y datos simulados generados mediante el programa estadístico **R**.

CAPÍTULO 2
PRELIMINARES

2.1 Poblaciones ocultas o escasas

Intentar definir una población oculta o escasa nos lleva a una serie de preguntas complejas. Si la principal característica de este tipo de poblaciones es que está oculta, ¿cómo puede estarse seguro de que existen?. Si existen, ¿cómo y por qué permanecen ocultas?

El término "población oculta" se refiere a un subconjunto de una población cuyos miembros poseen ciertas características que hacen que no puedan ser fácilmente distinguidos, ubicados o enumerados. En este tipo de situaciones la tarea de seleccionar un marco de muestreo apropiado es difícil, ya sea porque los individuos con la característica de interés conforman una subpoblación muy pequeña comparada con la población donde está contenida, o porque la característica de interés es estigmatizante o lesiva a su intimidad, lo que hace que sus miembros y los vínculos entre ellos permanezcan ocultos a la población general. Puede ocurrir que el investigador acuda a centros o lugares particulares que sean frecuentados por estas personas para obtener la muestra inicial, por ejemplo, si el estudio está relacionado con personas adictas a drogas, los individuos podrían localizarse en centros de rehabilitación o tratamiento [6].

Por otra parte, el término "oculto" se refiere a poblaciones de personas que son estigmatizadas por las características que poseen, las cuales pueden ser socialmente inaceptables o se ven con reserva, como por ejemplo: personas sin hogar, enfermos mentales, prostitutas, criminales, enfermos terminales, drogadictos, personas que

tienen preferencia sexual por individuos de su mismo sexo, personas que han sido abusadas sexualmente, e inmigrantes entre otras [7].

En conclusión, este tipo de poblaciones se caracteriza porque no existe un marco muestral definido de ellas, se desconoce su tamaño y principalmente porque sus miembros presentan una fuerte preocupación por su privacía, debido a que algunos de ellos pueden haber incurrido en conductas ilegales o estigmatizadas. Desde el punto de vista de recopilación de datos, se plantea el problema de que dichos individuos se nieguen a cooperar o que sus respuestas no sean confiables con el propósito de proteger su intimidad [2].

Por otra parte, en el proceso para estimar parámetros de una subpoblación escasa, el problema principal radica en que los métodos tradicionales y probabilísticos de muestreo no son efectivos para obtener estimados con suficiente precisión, excepto que se tome una muestra muy grande, que podría coincidir casi con la población. Para mostrar esta dificultad, supongamos que hay una población de tamaño N=6000, con una subpoblación de interés para el investigador de 600 individuos y se quiere tomar una muestra de n=100. Esto implica que existen $\binom{6000}{100} \simeq 3.05373 * 10^{219}$ maneras de tomar esta muestra, así que la probabilidad de obtener una muestra representativa es muy baja; Frank [1] expone que cuando la proporción de la subpoblación p, no se obtiene con suficiente precisión a menos que se tome una muestra relativamente grande comparada con el tamaño de la población, el tamaño de muestra para un estimador de p con una precisión relativa de 0.1 sería aproximadamente de $400/p$. En el ejemplo anterior se necesitaría una muestra de 4,000 elementos de la población de 6,000 para estimar p con una margen de error del 10%, lo que ya de por sí es muy elevado. El otro problema es la localización de la subpoblación debido precisamente a su condición de oculta o escasa. Otros problemas asociados a estudios con subpoblaciones ocultas o escasas son: la tasa de no respuesta y la veracidad de la respuesta, puesto que no generan muestras confiables, debido a que

la mayoría de poblaciones ocultas también son escasas [8].

Es importante mencionar que cuando se estudian poblaciones ocultas conviene aplicar muestreo por refencia o por redes como la forma más eficiente de obtener información por medio de cuestionarios elaborados especialmente.

2.2 Tipos de diseño de muestreo adaptativos

Cuando se tiene la intención y objetivo de realizar una investigación con esta clase de poblaciones la inquietud más importante que surge es tener una muestra representativa. Por lo expuesto anteriormente, las técnicas de muestreo probabilísticas no son las más indicadas. En el diseño de encuestas o estudios convencionales uno de los primeros procesos es la determinación del tamaño de la muestra, pues en estos casos el marco muestral se puede determinar con relativa facilidad. Algunas técnicas de muestreo probabilístico son: muestreo aleatorio simple, muestreo aleatorio estratificado y muestreo por conglomerado. En cambio, por la naturaleza de las poblaciones ocultas, el tamaño de la muestra prácticamente debiera coincidir con el universo para poder obtener una muestra representativa de la subpoblación, por lo cual se hace difícil llevar a cabo el estudio. Entre las técnicas no probabilísticas conocidas se encuentran: el muestreo por conveniencia, el muestreo por juicio, el muestreo por redes o "Rastreo por vínculos" entre otros, como los más importantes. En el proceso de estimación de parámetros de poblaciones ocultas, las técnicas y metodologías convencionales no son las más apropiadas a causa de la dificultad de localizar a los miembros con la característica deseada, y los diseños convencionales de muestreo probabilístico son a menudo poco prácticos a raíz de que no existe un marco de muestreo definido, por lo cual las inferencias que se hicieren pudieran no ser válidas. Además de esta dificultad también se desconoce el tamaño de la subpoblación y la identificación y la disposición del grupo específico. Para hacer frente a estas dificultades existe una técnica que agrupa todas las formas de recolección

de datos por medio de referencia, denominada Muestreo Adaptativo. Este tipo de muestreo se define como el proceso de selección por referencia de los elementos de la muestra dependiendo de los valores de las variables de interés que se observan durante el estudio [10]. Por ejemplo, la muestra final que se obtiene por medio de los contactos de un grupo de usuarios de droga inyectada (IV) en un estudio de una población determinada.

En un estudio realizado en el año 1994 por National Health and Social Life Survey [11] denominado Prácticas Sexuales en Estados Unidos, con respecto al síndrome de inmunodeficiencia adquirida (SIDA), se encontró que en una muestra de 3,422, personas sólo seis dieron positivo a la prueba, lo que refleja que los métodos convencionales son inadecuados. Por años, investigadores en comportamiento y ciencias biomédicas han percibido que los muestreos convencionales no son la mejor forma para recoger datos en poblaciones ocultas o raras. El uso de la información recolectada durante el proceso de muestreo es la diferencia dominante del muestreo adaptativo con respecto al muestreo convencional. En un muestreo convencional no es posible añadir o modificar la muestra después de realizado el proceso, aún si se descubre que hay un individuo más que es elegible. En cambio, en un estudio usando muestreo adaptativo, sobre la dependencia de la nicontina en jóvenes, en principio se aplica una entrevista a una muestra seleccionada al azar y en el transcurso de ella, la muestra se incrementa por medio de la red social del entrevistado, lo que no se podría hacer en el muestreo convencional. En conclusión, la selección de las personas a incluir en la muestra está basada en las observaciones hechas durante el estudio.

En los últimos años el muestreo adaptativo ha tenido un avance significativo en la teoría estadística. Los procedimientos estadísticos derivados de los cálculos para estimar parámetros de datos obtenidos por medio de esta técnica, son en algunos casos más precisos en términos del error cuadrático medio (MSE), en comparación

con estimadores basados en diseños convencionales [3]. Es importante reconocer que la teoría estadística de esta técnica es relativamente nueva.

En pocas palabras, cualquier diseño de muestreo que se adapta a las observaciones hechas en el transcurso del estudio es muestreo adaptativo. El muestreo adaptativo se clasifica en dos ramas generales: la denominada "Rastreo por Vínculos", la cual opera primeramente sobre una red social de información y el otro diseño que no tiene nombre propio, pero que opera básicamente usando información geográfica.

2.2.1 El diseño "Rastreo por vínculos"

El método de muestreo por "rastreo de vínculos" es una forma de recolectar información a través de las relaciones o vínculos sociales en la población de estudio. El "Rastreo por vínculos" clásico es el "snowball sampling" (muestreo por bola de nieve) propuesto por Leo Goodman [12] en 1961, que más adelante fue modificado por Ove Frank con conceptos de teoría de grafos [13]. Además está el "random walk" o camino aleatorio, diseñado por Klovdahl [14] y el "adaptive cluster sampling" (muestreo adaptativo por conglomerados) diseñado por Thompson [10]. Estos métodos se usan principalmente para formar una muestra de individuos que son referidos por otros elementos en la muestra para describir la población en términos de características e información individuales. En esta técnica se comienza con una muestra inicial ya sea obtenida a la azar, por una prueba piloto, o por medio del registro de alguna institución relacionada con el estudio a realizar. Por otro lado, los datos obtenidos a través de este método pueden también ser usados para describir la características estructurales de la población como complemento de los resultados de inferencia obtenidos por métodos convencionales [15].

Los investigadores de este tipo de poblaciones usan vínculos entre personas para encontrar otras que estén relacionadas e incluirlas en la muestra. El "Rastreo por

vínculos" es netamente adaptable porque el proceso de selección de la muestra depende de la relación social existente. Otros trabajos donde se han empleado estas técnicas son los realizados por J.F French (1993) [16] en Philadelfia y Newark acerca del intercambio de sexo por droga, y los de Biernacki (1986) quien investigó cómo usuarios de opio han logrado superarse por sí mismos de su adicción. Otros estudios incluyen consumo de cocaína y comportamientos sexuales [17], consumo de heroína [18, 19] y otras conductas más como prostitución y enfermedades de trasmisión sexual.(Baily and Aunger, 1995). La Universidad de Puerto Rico, Recinto de Río Piedras realizó un estudio de características socio-económicas de la población homosexual en el Área Metropolitana [20], después de hacer uso de diferentes medios de comunicación concluyeron que el método más eficiente en obtener más contactos fue "bola de nieve". A continuación se presenta una lista de los diferentes métodos de muestreo tipo "Rastreo por vínculos".

Muestreo por bola de nieve: Fue propuesto por Goodman [12] en el año 1961 y parte de la premisa de que todos los miembros de la población oculta pueden conocerce entre sí. Consiste en seleccionar una muestra inicial de individuos e identificar en cada entrevista qué otras personas nuevas de la población bajo estudio han de entrevistarse, para así ir formando la muestra. Generalmente la primera selección se hace en forma probabilística, mientras que las siguientes entrevistas quedan determinadas por las anteriores. En un sentido muy amplio, la primera muestra puede seleccionarse en forma intencional o estar constituida por voluntarios. Según Goodman, se parte de un grupo inicial de miembros y se les pide que mencionen a K personas, quienes siguen en turno con el mismo procedimiento hasta llegar a un número determinado de etapas u ondas.

Muestreo dirigido (Targeted Sampling): Combina el conocimiento de la ubicación geográfica de la población y técnicas de muestreo convencionales tales

como muestreo estratificado junto con métodos de rastreo. Inicialmente se identifican los estratos y luego se aplica el método adaptativo en cada uno de ellos [21].

Muestreo por Multiplicidad: Este tipo de muestreo fue introducido por Birnbaum y Sirken [22] para estimar en particular el número de pacientes de fibrosis cística. Se decidió aplicar este tipo de muestreo debido a que muchos centros de salud proveen tratamiento y atención a personas que la padecen, y que pueden ser reportadas varias veces por ellos. En sí, una persona enferma, que ha sido tratada en más de un centro médico tiene una probabilidad mayor de ser incluida en la muestra por referencia que una persona que sólo ha sido tratada en un sólo centro médico. Por lo tanto, tal situación debe ser tomada en cuenta para la estimación de los parámetros respectivos. Una variedad de reglas de vínculo y diseños de muestreo han sido investigados bajo la teoría general de muestreo por redes ("network sampling") [9, 23, 24].

"Random Walk": También conocido como "camino aleatorio". Fue propuesto por Klovdahl(1989) [14] y consiste en una variación del muestreo por rastreo. El procedimiento consiste en escoger una persona al azar y se le pregunta que refiera a otros miembros de la población oculta. De esta lista se selecciona uno al azar y se añade a la muestra y se le pide que haga nuevamente una lista y así en adelante, hasta conseguir un número determinado de personas. La motivación de este procedimiento es que la muestra inicial de individuos debe se atípica en sus características.

"Respondent-Driven Sampling (RDS)": Muestreo dirigido por encuestados. Este método es similar al de "bola de nieve" en el sentido de que involucra cadenas de referencia, sin embargo el proceso de reclutamiento de una manera permite el uso de probabilidades. Se parte de un grupo inicial temporal de miembros de la población quienes contactan un número fijo de miembros por medio de un código único de registro, según las características que demande el estudio. La participación en el estudio es totalmente voluntaria, así que el nuevo miembro contactado debe

presentarse al lugar del estudio y dar su código; en ningún momento el encuestador necesita la identificación de los contactos. En esta técnica ambas partes reciben un estímulo económico. Es importante anotar que el número de contactos es fijo, contrario al de "bola de nieve", además los contactos son registrados y el tamaño de la red es controlado. Entre las ventajas de este método se tiene que no es necesario un marco de muestreo y que la tasa de no respuesta disminuye debido al incentivo económico y a la partipación voluntaria. La diferencia principal entre el RDS y el de "bola de nieve" es que el grupo inicial tiene un número limitado de contactos, por lo que se minimiza la influencia de la muestra inicial sobre la final. Otra diferencia es que se registra la relación entre el grupo inicial y las referencias sucesivas que van haciendo los elementos que aparecen en la muestra, así que la multiplicidad de los contactos puede ser evaluada y ajustada para el análisis.

2.2.2 Otros tipos de muestreo adaptativo

Otros tipos de muestreo adaptativo son aquellos que toman en cuenta información geográfica donde se ubican los individuos objetos de estudio. Entre ellos se mencionan:

Conglomerado Adaptativo: En este tipo de muestreo los vínculos de los miembros de la muestra son agregados si la variable de interés se satisface [10]. En forma sencilla, una muestra inicial se toma por medio de una técnica de muestreo convencional, si un individuo en la muestra satisface la característica, entonces todos los individuos vinculados a él independientemente que tengan la característica o no, en una área específica son añadidos a la muestra, y el proceso continúa de la misma manera, con estos nuevos contactos. Por ejemplo, en un estudio de dependencia de los jóvenes a la nicotina se toma una muestra en vecindades, si alguna residencia tiene un joven fumador, entonces se deben obtener muestras en las casas cercanas del vecindario y así en adelante. En este tipo de muestreo, los vecinos se definen tanto social como geográficamente, es decir que el significado de vecino es más amplio.

Ubicación y estratificación adaptativa: Se refiere al diseño en el cual la subpoblación se divide en estratos claramente definidos por cierta clasificación y las observaciones se ubican dentro de dichos estratos.

Ubicación Adaptativa: Se obtiene primero una muestra por medio de un muestreo al azar o por medio de estratos. Esta muestra es seleccionada para verificar cuál o cuáles áreas tienen más evidencia de las características de interés. Entonces el muestreo se concentra en esas áreas representativas. Otras estrategias de ubicación adaptativa han sido descritas por Solomon y Zacks [25].

2.2.3 La teoría de grafos en "Rastreo por vínculos"

Para entender acertadamente los contactos y referencias es necesario realizar una descripción completa y rigurosa de la estructura de sus relaciones a través de un modelo que permita un análisis eficiente. El uso de conocimientos matemáticos y de teoría de grafos en el estudio de redes sociales nos permite representar las características de una red de manera concisa y sistemática. Los modelos de redes son ampliamente utilizados para representar las relaciones, las interacciones e implicaciones entre ellas [26].

En el campo de análisis de redes, la teoría de muestreo se ha asociado con términos y propiedades de grafos aleatorios (random graph) por medio de trabajos realizados por Erdös y Rényi [27]. Durante las pasadas décadas se han hecho esfuerzos para descubrir propiedades de las redes sociales con ayuda de la teoría de grafos. Los investigadores han desarrollado métodos para estimar la media y varianza del grado de un grafo [13]; la probabilidad de que un grafo sea conectado [28]; la distribución de los componentes conectados y otras estimaciones en grafos de gran tamaño bajo varios marcos de muestreo [5].

Las redes pueden describir el comportamiento de epidemias, datos genéticos, movimientos en aeropuertos, transmisión de señales, entre otras. Para efectos de nuestros estudio, una red social consiste de un conjunto de n nodos y de todas las posibles relaciones existentes entre ellos, representados por las parejas (i, j), donde $i, j \in \{1, ..., n\}$. Los datos a menudo se representan por medio de la matriz de adyacencia $n \times n$. La presencia o ausencia de un nodo en la relación se indica por 1 ó 0 respectivamente. En el caso de relaciones binarias, los datos también pueden ser representados por medio de un grafo con sus nodos y arcos. La dirección de los arcos de un grafo, dada la situación a representar puede ser independiente de un vértice a otro (dirigido), es decir $(i, j) \neq (j, i)$ o puede ser un arco mutuo (no dirigido), $(i, j) = (j, i)$.

El análisis de redes sociales junto con el diseño del muestro "Rastreo por vínculos" ha contribuido notablemente a la estimación e inferencia de parámetros en poblaciones ocultas o escasas. La combinación de los conceptos teóricos junto con los procesos computacionales permiten que se fortalezca y consolide como un método de muestro más eficiente [26] que los convencionales en estos casos.

CAPÍTULO 3
TRABAJOS PREVIOS

Ove Frank y Tom Snijders en un artículo publicado en el año 1994 [1] plantearon el desarrollo teórico y práctico del muestreo "bola de nieve" para estimar el tamaño de poblaciones ocultas y socialmente sensitivas. Ellos desarrollaron dos tipos de estimadores denominados **"Estimación con base en el Modelo"** y **"Estimación con base en el Diseño"**; el primero, parte de que la muestra inicial de nodos y el conjunto de arcos tiene una distribución de probabilidad específica, a partir de la cual se identifica y cuantifica el número de individuos con sus respectivos contactos y demás estadísticos, para luego obtener diferentes estimadores ; y el segundo se enfoca en encontrar los parámetros a partir de un grafo fijo, dirigido y desconocido, donde los vértices y los arcos no son variables aleatorias. De nuevo se asume que la muestra inicial tiene una distribución particular y que los diversos estadistícos se obtienen a partir de valores esperados. A continuación se resumen las ideas principales de O. Frank y T. Snijders sobre estimación de parámetros en poblaciones ocultas con algunos conceptos de teoría de grafos.

La población oculta, junto con los contactos entre sí, se considera naturalmente un grafo que podría ser dirigido o no dirigido. Los miembros de la población son los vértices del grafo. El número de vértices, v, se estima de la información de la muestra. El vértice i tiene un arco a j, si el individuo i refiere al individuo j cuando le preguntan por sus contactos. Se supone que una muestra de n miembros de la población oculta está disponible. Cada uno de estos miembros mencionan otros

miembros que conozcan o pertenezcan a la población. Puede pasar que algunos de ellos sean mencionados por varios individuos y estén dentro o fuera de la muestra inicial S_0. Aquellos que estén fuera de S_0 decimos que pertenecen a la primera onda, S_1. Los contactos que no estén ni en S_0 ni en S_1, y que son mencionados al menos por un miembro de S_1, pertenecen a S_2 y así sucesivamente. La muestra final consiste de la muestra inicial junto con todas las ondas generadas de ella. Teóricamente se dice que el proceso termina cuando la última onda no proporciona miembros diferentes a los mencionados en las ondas anteriores. La muestra inicial puede ser seleccionada por muestreo al azar simple de una población que contiene subpoblaciones ocultas o raras. La selección de cada individuo para la muestra inicial es independiente y con distribución Bernoulli, dependiendo si tiene o no la característica de interés, con parámetro α. La mayoría de las veces una muestra aleatoria no proporcionaría miembros de la población oculta o proporcionaría muy pocos, por lo que el investigador se vería obligado a acudir a lugares característicos, típicos o muy frecuentados dependiendo del tipo de variable que se desea estudiar, por ejemplo si se desea conocer aspectos relacionados con el consumo de droga se podrían visitar bares, clubes nocturnos o centro de rehabilitación, entre otros.

Conceptos del muestreo por "bola de nieve"

Como se mencionó anteriormente, la población se considera un grafo dirigido de V vértices, es decir

$$V=\{1,2,3,...,v\}$$

Los arcos son pares ordenados (i,j) de vértices de V. El conjunto de arcos W entre los vértices es un subconjunto de V^2. La muestra inicial y el conjunto de arcos se representan respectivamente por las variables:

$$x = (x_i : i \in V)$$

$$y = (y_{ij} : (i,j) \in V^2)$$

- x_i es 1 ó 0 de acuerdo si el vértice i está o no en la muestra inicial.
- y_{ij} es 1 ó 0 si el grafo contiene o no un arco de i a j.

Al conjunto y se le denomina la matriz de adyacencia del grafo. A continuación presentamos las principales notaciones del modelo.

- $A_j = \{i \in V : y_{ji} = 1\}$: Representa un subconjunto de vértices después del vértice j. A_j es indicada por la fila j de la matriz y.
- $B_j = \{i \in V : y_{ij} = 1\}$: Representa un subconjunto de vértices antes del vértice j. B_j es indicado por la columna j de la matriz y.
- $|A_j| = a_j = \sum(y_{ji})$: Grado de salida del vértice j.
- $|B_j| = b_j = \sum(y_{ij})$: Grado de entrada del vértice j.

Para cualquier $S \subset V$ se define:

$A(S) = \cup_{j \in S}(A_j)$: Vértices después de cada vértice en S

$B(S) = \cup_{j \in S}(B_j)$: Vértices antes de cada vértice en S

S_0: Muestra inicial.

$S_1 = A(S_0) \cap \overline{S_0}$= Primera Onda.

$S_2 = A(S_1) \cap \overline{S_0} \cap \overline{S_1}$= Segunda onda.

$S_f = S_0 \cup S_1 \cup S_2 ... \cup S_k$,

donde, S_f es la muestra final, k número de etapas y $S_{k+1} = \varnothing$.

Estimación según el modelo

O. Frank y T.Snijders [1] plantean que la muestra inicial S_0 se distribuye Bernoulli con probabilidad α. Esto significa que cada uno de los vértices o miembros $x_1, x_2, ..., x_v$ tiene una probabilidad α, de tener la característica. Debido a que la sumatoria de n distribuciones Bernoulli independientes tiene una distribución binomial, el tamaño de la muestra inicial $n = |S_0|$ se distribuye Binomial(v,α). Además se supone que el conjunto de arcos W se distribuye Bernoulli con probabilidad β. En el grafo dirigido se consideran los siguientes estadísticos.

r: Número de arcos en S_0.

Condicionalmente sobre n, la distribución de r y su esperanza están dadas por:

$$r \sim Bin(n(n-1), \beta)$$

por lo que

$$E(r) = n(n-1)\beta$$

s: Número de arcos de S_0 a S_1,

$$s = |W \cap (S_0 \times S_1)|$$

Condicionalmente sobre n, la distribución y la esperanza de s están dadas por:

$$s \sim Bin(n(v-n), \beta)$$

y por lo tanto

$$E(s) = n(v-n)\beta.$$

Los estimadores de momento para r y s, dependiendo de n se obtienen a partir de la función generatriz de momentos de la distribución binomial (n, p)

$$m(t) = [pe^t + (1-p)]^n$$

cuando se evalua su derivada en $t = 0$, entonces

$$r = n(n-1)\beta$$

$$s = n(v-n)\beta$$

y por consiguiente los estimados respectivos son:

$\widehat{\beta}_1 = \frac{r}{n(n-1)}$

$\widehat{v}_1 = \frac{nr+(n-1)s}{r}$.

Un segundo estimador para v puede obtenerse a partir de la siguiente información:

la frecuencia de los arcos r, s son independientes sobre n, cuya suma $t = s + r$ se distribuye binomial$(n(v-1), \beta)$, además el tamaño de $|S_1| = m$ de la primera onda es tambien binomial$(v-n, 1-(1-\beta)^n)$. Las dos ecuaciones derivadas por la función generatriz de momentos de sus respectivas distribuciones son las siguientes:

$$t = n(v-1)\beta$$

y

$$m = (v - n[1 - (1-\beta)^n])$$

donde el estimador de v es la solución a la siguiente ecuación, consecuencia de las dos ecuaciones anteriores:

$$\frac{v-n-m}{v-n} = \left[\frac{n(v-1)-t}{n(v-1)}\right]^n$$

La solución a esta ecuación se puede denotar como $\widehat{v}_2$. Existe otra forma para estimar $\hat{v}_2$. Sea m_k el número de individuos que no están en S_0, que son mencionados por exactamente k miembros de S_0. Entonces

$m = m_1 + m_2 + \cdots + m_n,$

$s = m_1 + 2m_2 + \cdots + nm_n,$

$m_0 = v - n - m,$

donde m es el número de contactos inmediatos después de la muestra inicial S_0 y

$$(m_1, m_2 \ldots m_n) \sim multinomial(v-n, p_1, p_2 \ldots p_n),$$

donde $p_k = \binom{n}{k}\beta^k(1-\beta)^{n-k}$.

Como r y $(m_1, m_2 \ldots m_n)$ son estadísticos y condicionalmente independientes, y el producto de ellos está dado de la siguiente forma:

$$L(r, m|n) = \binom{n(n-1)}{r}\beta^r(1-\beta)^{n(n-1)-r}(v-n)!\prod_{k=0}^{n}(\frac{p_k^{m_k}}{m_k!})$$

que se puede reescribir de la siguiente forma:

$$L(r,m|n) = \frac{(v-n)!}{(v-n-m)!}\binom{n(n-1)}{r}\beta^t(1-\beta)^{n(v-1)-t}\prod_{k=1}^{n}\frac{\binom{n}{k}^{m_k}}{m_k!}$$

y como m y t son estadísticos suficientes, la parte principal teniendo en cuenta v y β de la función de verosimilitud es:

$$g(\beta,\alpha) = \frac{(v-n)!}{(v-n-m)!}\beta^t(1-\beta)^{n(v-1)-t}$$

fijando v y derivando se tiene

$$\frac{\partial g(\beta,v)}{\partial\beta} = t\beta^{t-1}(1-\beta)^{n(v-1)-t} - \beta^t(n(v-1)-t)(1-\beta)^{n(v-1)-t-1}$$

Igualando a cero y despejando se tiene que $g(\beta,v)$ se maximiza por $\beta = \frac{t}{n(v-1)}$. Ahora, si se fija β la derivada en este caso resulta un poco difícil por la presencia del término factorial. Frank [1] plantea resolver la ecuación $g(\beta,v) = g(\beta,v-1)$, cuya solución es equivalente a $(v-n-m) = (v-n)(1-\beta)^n$ y por lo tanto el valor del estimador $\widehat{v}_2$.

Para la estimación de las varianzas de cada estimador de los vértices, Frank y Snijders utilizaron series de Taylor de primer orden y obtuvieron los siguientes estimados

$$var(\widehat{v}_1|n) \simeq \frac{(v-1)(v-n)(1-\beta)}{\beta n(1-n)}$$

que se calcula por el reemplazo de $\widehat{v}_1$ y por $\widehat{\beta}_1$, por lo tanto

$$var(\widehat{v}_1) = \frac{(n^2-n-r)(n-1)s(s+r)}{nr^3}.$$

De manera similar la varianza estimada de $\widehat{v}_2$ está dada en términos de $\hat{v}_2$ debido a que no es posible expresarlo en forma algebraica.

$$var(\widehat{v_2}) = \frac{(\hat{v}_2 - n)^2}{(t-m)}.$$

De acuerdo con las varianzas estimadas, los intervalos de confianza de un 95% para v asumiendo normalidad para los estimadores son:

$$\widehat{v_1} \pm 1.96\sqrt{var(\widehat{v_1})},$$

en el caso de $\widehat{v}_1$ y sustituyendo por los respectivos estimados

$$\widehat{v_1} \pm 1.96\sqrt{\frac{(n^2-n-r)(n-1)(s-r)s}{nr^3}}$$

y cuando se usa el segundo el estimador de v su intervalo de confianza está dado por

$$\widehat{v_2} \pm 1.96\sqrt{var(\widehat{v_2})}$$

y reemplazando se tiene que

$$\widehat{v_2} \pm 1.96\frac{(\widehat{v_2} - n)}{\sqrt{t-m}}$$

Estimación según el diseño

Este diseño, presentado por Frank y Snijders [1], tiene como fin estimar el número de vértices de un grafo fijo y desconocido. Aquí nuestra población se representa por un grafo, como se mencionó anteriormente. De nuevo, los estadísticos n, r, s, m son considerados. Puesto que $x_1, x_2, ..., x_n$ se distribuyen independientemente Bernoulli con parámetro α, el valor esperado de cada estadístico se presenta en las siguientes expresiones en términos de vértices y arcos:

$E(n) = E\sum_i x_i = v\alpha$

$E(r) = E\sum_{i\neq j} x_i x_j y_{ij} = (w-v)\alpha^2$, donde $w = |W|$

$E(s) = E\sum_{i\neq j} x_i(1-x_j)y_{ij} = (w-v)\alpha(1-\alpha)$

$E(m) = E\sum_j(\max_{i\in B_j} x_i - x_j) = E\sum_j(1 - x_j - \min_{i\in B_j}(1-x_j))$

$E(m) = v(1-\alpha) - \sum_j(1-\alpha)^{b_j}$

De los tres primeros valores esperados se pueden obtener los siguientes estimadores muy similares a los expuestos anteriormente para $\alpha, v,$:

$\widehat{\alpha}_3 = r/(r+s)$

$\widehat{v}_3 = n(r+s)/r$

Además el estadístico m se puede combinar con el estadístico k, definido como el número de vértices en la muestra inicial conectados con al menos un vértice de la muestra inicial, para obtener otro estimador. Simbólicamente k se escribe

$$k = \sum_j x_j \max_{i\neq j} y_{ij} x_i$$

cuya esperanza es $E(k) = v\alpha - \alpha\sum_j(1-\alpha)^{b_j-1}$. Con propiedades del valor esperado y procedimientos algebraicos se tiene que $E(k) = \alpha E(k+m)$; por consiguiente el estimador derivado es:

$\widehat{v}_4 = \frac{n(k+m)}{k}$ con su respectivo $\widehat{\alpha}_4 = \frac{k}{k+m}$

En el proceso de simulación realizados por los autores con respectivos grafos se verificó que el mejor estimador en términos de varianza fue $\hat{v}_2$, y teniendo en cuenta la exigencia de la información necesaria para el cálculo de éstos, concluyeron que se pueden clasificar en una escala de menor a mayor como lo indica su respectivo subíndice.

CAPÍTULO 4
INFERENCIA DEL DISEÑO ESTRATIFICADO CON RASTREO DE VÍNCULOS

En estudios con poblaciones ocultas y de difícil acceso, el método "Rastreo por vínculos" se convierte en la mejor y más práctica forma de obtener una muestra representativa para realizar estimaciones sobre la población [10]. La inferencia de la muestra sobre la población puede ser afectada por la forma en que se toma la muestra y por la información contenida en ella, por ello es importante tener en cuenta la naturaleza y comportamiento de la población en el momento de recolectar los datos. En términos de relaciones y grafos, la población puede ser vista esencialmente como una estructura fija o como un grafo completamente estocástico, que refleja y describe características de los individuos y las relaciones existentes entre cada uno de ellos. En cada uno de los tipos de muestreo relacionados con "Rastreo por vínculos" como "snowball sampling", "random walk, "network sampling" entre otros, se supone que el proceso de selección de la muestra inicial tiene una distribución Bernoulli. En particular, el diseño "bola de nieve", avanzó rápidamente gracias a los trabajos de Frank por medio de la teoría de grafos [1]. Precisamente, como se habló en el capítulo pasado Frank y Snijders en 1994 plantearon la estimación del tamaño de la población oculta a partir de los vértices que representan cada individuo de interés en la población de estudio.

4.1 Modelo del grafo y diseño del muestreo

Con este nuevo enfoque Frank y Thompson consideran un grafo con N nodos, reunidos en un conjunto denominado U, es decir

$$U = \{1, 2, 3, ..., N\},$$

donde cada uno de los nodos está asociado con una variable de interés Y_n que indica la presencia o no del nodo, que a la vez están reunidas en un conjunto llamado $\mathbf{Y}$ y dado por $\mathbf{Y} = (Y_1, Y_2, ..., Y_N)$. Para cualquier par de nodos u y v, que pertenezcan a U se define la variable X_{uv}, que indica la presencia o no de arco del vértice u al vértice v. Es decir $X_{uv} = 1$, si existe arco de u a v (arco direccional) y $X_{uv} = 0$ de otra forma. El conjunto de todos los arcos define la matrix de adyacencia que se simboliza por $\mathbf{X}$. Los elementos de la diagonal de la matriz $\mathbf{X}$ son ceros, pues no se considera que un miembro de la población se referencie a sí mismo. Es importante mostrar que el par ordenado (u, v) está asociado al valor de las variables $(Y_u, Y_v, X_{u,v}, Y_{v,u})$ agrupadas convenientemente en este tetra ordenado.

El esfuerzo por recolectar datos a través de "Rastreo de vínculos" requiere de gran atención cuando la población es de un tamaño considerable debido al proceso de estimar propiedades de las redes de una muestra seleccionada. Como se ha mencionado anteriormente, las relaciones entre la población se considera un grafo con sus nodos y arcos respectivos. El modelo del grafo está dado por la función de probabilidad

$$f(y, x, \psi),$$

donde $y \in \mathbf{Y}, x \in \mathbf{X}$, nodos y arcos respectivamente y ψ los parámetros desconocidos. La muestra de nodos y de pares de nodos del grafo se representa por s. La muestra s en sí está compuesta por dos subconjuntos así, $s = (s^1, s^2)$, donde s^1 representa el subconjunto de nodos y s^2 representa el subconjunto de pares de nodos. La información recolectada en la muestra s con el valor de las variables definidas

anteriormente se denota por $d = (s, y_{s^1}, x_{s^2})$, sin embargo, es a menudo conveniente escribir $d = (s, y_s, x_s)$ que indica los valores y de los nodos en la muestra combinada al igual que los valores x de los pares nodos, junto con $y_{\overline{s}}$ y $x_{\overline{s}}$ que representan el valor de los nodos y pares de nodos que no fueron seleccionados. El diseño del tipo de muestreo puede estar controlado por el investigador, o puede estar más allá del control de él y sujeto a condiciones de la población. Si la probabilidad de selección de la muestra no depende del valor de los nodos ni de los pares de nodos o de algún parámetro ψ relacionado con el modelo del grafo, decimos que el diseño es convencional, y la probabilidad de selección puede escribirse como $p(s)$ o $p(s; \varphi)$, donde φ denota cualquier parámetro involucrado con el diseño más no con el modelo, por ejemplo en el muestreo Bernoulli donde no se conoce la probabilidad de inclusión φ de cada nodo. En cambio, si la probabilidad de muestreo depende sobre algún valor y o x, se denomina diseño "adaptativo", pues el proceso de selección de la muestra se adapta a la configuración de los nodos y vínculos en la población. En general, el diseño de muestreo en el contexto de la teoría de grafos tiene una probabilidad de selección que se puede escribir como $p(s|y, x; \psi)$ donde y denota la secuencia del valor de los nodos, x la matriz del valor de los arcos, y ψ algún parámetro.

La muestra $d = (s, y_s, x_s)$ es una función de la muestra seleccionada y de los valores y y x en el grafo, por lo tanto la función que representa al grafo teniendo en cuenta, el modelo y diseño es:

$$L(\psi, d) = \sum p(s|y, x; \psi) f(y, x; \psi),$$

cuya suma es sobre todos los valores y, x consistentes con la datos d. Puesto que los valores y y x de los nodos y pares de nodos en la muestra son fijos en el grafo, la suma es sobre todos los posibles valores no observados de las variables $y_{\overline{s}}$ y $x_{\overline{s}}$. Por lo tanto, cualquier diseño en el cual la selección de la muestra dependa sobre los valores y, x del grafo sólo por medio de y_s y x_s incluidos en la muestra, la

probabilidad del diseño puede ser tomado como una constante en la sumatoria de la función y obviamente como un factor que multiplica la sumatoria. En conclusión la función $L(\varphi, \psi, d)$ se podría reescribir así:

$$L(\varphi, \psi, d) = p(s|y_s, x_s; \varphi) \sum_{y_{\bar{s}}, x_{\bar{s}}} f(y, x; \psi),$$

donde φ denota los parámetros del diseño y ψ los parámetros del modelo.

4.2 Formulación teórica de Frank, Thompson y Chow

En esta sección y en la siguiente se presentan los desarrollos teóricos realizados por Frank, Thompson y Chow [4, 10], en lo que respecta al enfoque bayesiano y los estimadores basados en la función del modelo y diseño del grafo.

4.2.1 Relación de los vínculos con el valor de los nodos.

Esta clase de modelo se construye con una independencia condicional entre las tetras ordenadas $(Y_u, Y_v, X_{u,v}, Y_{v,u})$ al igual que entre los contactos del modelo. Es decir, condicionalmente sobre el valor de los nodos se asume independencia de las tetras, donde la distribución de los vínculos entre los pares de nodos depende del valor de ellos.

En este modelo se asume que las variables asignadas a cada nodo $Y_1, ..., Y_N$ son independientes e identicamente distribuidas Bernoulli con probabilidad de selección $P(Y_u = i) = \theta_i$, para $i = 0, 1$, con $\theta_0 + \theta_1 = 1$. Condicionalmente sobre el valor de las variables de los nodos $Y_1, ..., Y_N$, las parejas (X_{uv}, X_{vu}) son independientes para $1 \leq u < v \leq N$ y con distribución

$$P[(X_{uv}, X_{vu}) = (k, l)|Y_u = i, Y_v = j] = \lambda_{ijkl}$$

para todas las combinaciones $i = 0, 1; j = 0, 1; k = 0, 1; l = 0, 1$. Para simplificar la suma de la probabilidades, se opta por escribir un punto en lugar del índice para mostrar la sumatoria sobre dicho índice. Por ejemplo, para todas las combinaciones

sobre i, j la suma sobre k, l está denotada por $\lambda_{ij\bullet\bullet} = \sum_k \sum_l \lambda_{ijkl}$.

Es importante saber que para obtener el valor de las probabilidades de los vínculos, sin depender de identificación de nodos, las siguientes identidades son deducidas: $\lambda_{1110} = \lambda_{1101}, \lambda_{1011} = \lambda_{0111}, \lambda_{1010} = \lambda_{0101}, \lambda_{1001} = \lambda_{0110}, \lambda_{0010} = \lambda_{0001}, \lambda_{1000} = \lambda_{0100}$, Por ejemplo, $\lambda_{1110} = \lambda_{1101}$ explica la igualdad de la probabilidad existente entre dos vértices con la característica y con una arco direccional. Por otra parte, es conveniente denotar $\lambda_{ij1\bullet} = \sum_l \lambda_{ij1l} = \alpha_{ij}$ y $\lambda_{ij11} = \beta_{i+j}$, para $i = 0, 1$ y $j = 0, 1$.

El valor α_{ij} es la probabilidad de arco de un nodo de valor i a un nodo de valor j.

El valor β_k es la probabilidad de arco mutuo entre k nodos marcados.

N_i denota el número total de nodos con valor i en el grafo, para $i = 0, 1$ así que $N_0 + N_1 = N$.

Además M_{ijkl} denota el total de tetras ordenadas de tipo $(ijkl)$, de pares ordenados (u, v), con $u < v$, tal que $(Y_u, Y_v, X_{uv}, X_{vu}) = (ijkl)$, por lo cual la función del grafo completo bajo el modelo con parámetros (θ, λ) es

$$L(\theta, \lambda; y, x) = \left(\prod_{i=0}^{1} \theta_i^{N_i} \right) \left(\prod_{i=0}^{1} \prod_{j=0}^{1} \prod_{k=0}^{1} \prod_{l=0}^{1} \right) \lambda_{ijkl}^{M_{ijkl}},$$

y para indicar la cantidad de tetras ordenadas $(ijkl)$ se emplea la siguiente notación: R_{ab}, donde a es la suma del valor de los nodos y b es la suma del valor de los arcos, aunque es necesario aclarar que si $a = 1$ y $b = 1$ se tiene dos divisiones: $\tilde{R}_{11} = M_{0101} + M_{1010}$ y $\grave{R}_{11} = M_{0110} + M_{1001}$ que indican el número de tetras ordenadas $(ijkl)$ con un arco de un nodo marcado a un nodo no marcado, y de uno no marcado a uno marcado respectivamente, para las demás R_{ab} no existe ninguna condición.

4.2.2 Inferencia del diseño Rastreo de vínculos

Para cualquier diseño de "rastreo de vínculos" con muestra inicial S_0, con nodos adyacentes, S_1, con sus respectivos vínculos o arcos y $d = (s, y_s, x_{s_0U})$, con el

modelo del grafo planteado anteriormente, la función de $L(\theta, \lambda; d)$ con la muestra seleccionada es:

$$L(\theta, \lambda; d) = p(s|y_s, x_{s_0U}) \sum \left(\prod_{u=1}^{N} \theta_{y_u} \right) \left(\prod_{u<v} \lambda_{y_u y_v x_{uv} x_{vu}} \right),$$

donde la suma es sobre todos los valores y_u y x_{uv} que no están fijos en la muestra. Para cuantificar y denotar el tipo de nodo en un conjunto A cualquiera se tiene en cuenta la siguiente notación: $n_i(A)$ indica el número de nodos $u \in A$ con $y_u = i$, para cualquier subconjunto $A \subset U$, y m_{ijkl} el número de pares ordenados (u, v) tal que $u \in A$ y $v \in B$ que cumplen $(y_u, y_v, x_{uv}, x_{vu}) = (ijkl)$. Por lo tanto la función $L(\theta, \lambda; d)$ se escribe de la siguiente forma:

$$p(s|y_s, x_{s_0U}) \prod_i \theta_i^{n_i(s)} \prod_{ijkl} \lambda_{ijkl}^{m_{ijkl}(s_0,s_0)} \prod_{ijk} \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0,s_1)} \prod_{v \in \bar{s}} \left[\sum_j \theta_j \prod_{ik} \lambda_{ijk\bullet}^{m_{i.k\bullet}(s_0,v)} \right]$$

Como todos los vínculos de la muestra inicial son trazados entonces la submatriz $\mathbf{x}_{s_0\bar{s}}$ es cero y $m_{i\bullet 0\bullet}(s_0, v) = n_i(s_0)$ para $v \in \bar{s}$, lo que implica que la función $L(\theta, \lambda; d)$ se simplifica aún más así:

$$L(\theta, \lambda; d) = p(s|y_s, x_{s_0U}) \prod_i \theta_i^{n_i(s)} \prod_{ijkl} \lambda_{ijkl}^{m_{ijkl}(s_0,s_0)} \prod_{ijk} \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0,s_1)} \left[\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)} \right]^{n(\bar{s})}$$

El factor $\prod \theta_i^{n_i(s)}$ da la probabilidad de observar el valor de los nodos en la muestra.
El factor $\prod \lambda_{ijkl}^{m_{ijkl}(s_0,s_0)}$ da la probabilidad de observar los tipos de parejas (tetras) dentro de (s_0, s_0).
El factor $\prod \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0,s_1)}$ da la probabilidad de observar parejas (tetras) en (s_0, s_1). Como sólo se observa el vínculo x_{uv}, $u \in s_0$, $v \in s_1$, la probabilidad de que $x_{uv} = k$ dado $y_u = i$ y $y_v = j$ es $\lambda_{ijk\bullet}$. El factor que se encuentra en medio de llaves indica la probabilidad de no arcos de la muestra inicial a $\bar{s}$. Y por último el factor θ_j indica

la probabilidad $y_v = j$ para un nodo que pertence a $n(\bar{s})$.

Teniendo en cuenta que consideramos un modelo simétrico, es decir que $\lambda_{ijkl} = 0$ para $k \neq l$, entonces la representación de las relaciones es una clase de grafo no dirigido. Luego $\lambda_{ij11} + \lambda_{ij00} = 1$ y como $\beta_{i+j} = \lambda_{ij11}$ entonces la función del modelo y el diseño se escribe:

$$L(\theta, \beta; d) = p(s|y_s, x_{s_0U}) \prod_i \theta_i^{n_i(s)} \prod_{ij} \beta_{i+j}^{m_{ij11}(s_0,s)} (1-\beta_{i+j})^{m_{ij00}(s_0,s)} \left[\sum_j \theta_j \prod_i (1-\beta_{i+j})^{n_i(s_0)} \right]^{n(\bar{s})}$$

Ahora para simplificar la expresión aún más y obtenerla de la forma más apropiada aplicamos la siguientes notaciones:

$$\begin{aligned} r_{0,0} &= m_{0000}(s_0, s) & r_{0,2} &= m_{0011}(s_0, s) \\ r_{1,0} &= m_{0100}(s_0, s) + m_{1000}(s_0, s) & r_{1,2} &= m_{0111}(s_0, s) + m_{1011}(s_0, s) \\ r_{2,0} &= m_{1100}(s_0, s) & r_{2,2} &= m_{1111}(s_0, s) \end{aligned}$$

Por lo tanto la función $L(\theta, \beta; d)$ se simplifica así:

$$\begin{aligned} L(\theta, \beta; d) = p(s|y_s, x_{s_0U}) &\theta_0^{n_0(s)} (1-\theta_0)^{n_1(s)} \beta_0^{r_{0,2}} (1-\beta_0)^{r_{0,0}} \beta_1^{r_{1,2}} (1-\beta_1)^{r_{1,0}} \beta_2^{r_{2,2}} (1-\beta_2)^{r_{2,0}} \\ &\left[\theta_0 (1-\beta_0)^{n_0(s_0)} (1-\beta_1)^{n_1(s_0)} + (1-\theta_0)(1-\beta_1)^{n_0(s_0)} (1-\beta_2)^{n_1(s_0)} \right]^{n(\bar{s})} \end{aligned}$$

Desde este punto consideramos la función $L(\theta, \beta; d)$ en el contexto bayeasino. En primera medida asumimos independencia entre las distribuciones a priori de cada uno de los parámetros $\theta_0, \beta_0, \beta_1, \beta_2, .$ En la teoría bayesiana el parámetro es considerado una variable que puede ser descrita por medio de una distribución de probabilidad llamada distribución a priori, que puede ser objetiva o derivada de un experimento y formulada antes de tomar la muestra. Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de una población con parámetro denotado por θ. Entonces la información

obtenida de la muestra junto con la distribución a priori conlleva a obtener una distribución llamada a posteriori. Escrito en lenguaje matemático, la distribución a priori se denota por $\pi(\theta)$, la distribución del muestreo por $f(x|\theta)$ y la distribución posterior por $\pi(\theta|x) = f(x|\theta)\pi(\theta)/m(x)$, donde $m(x)$ es la distribución marginal de la muestra $\mathbf{X}$, es decir

$$m(x) = \int f(x|\theta)\pi(\theta).$$

La distribución a posteriori también se puede escribir como

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$$

que indica que es proporcional al producto de la a priori y de la muestra, pues el denominador es una constante.

Puesto que no conocemos el valor de los parámetros $\theta_0, \beta_0, \beta_1, \beta_2$, y se sabe que los valores oscilan en el intervalo $[0, 1]$ se asume que cada uno tiene distribución $Beta$, debido a que es el tipo de familia que más modelos ofrece entre 0 y 1. En consecuencia, la función de distribución a priori para $L(\theta, \beta; d)$ es proporcional a la multiplicación de cada una de las distribuciones apriori con parámetros, $a, b, c, d, e, f, g, h > 0$.

$$\pi(\theta_0, \beta_0, \beta_1, \beta_2) \propto \theta_0^{a-1}(1-\theta_0)^{b-1}\beta_0^{c-1}(1-\beta_0)^{d-1}\beta_1^{e-1}(1-\beta_1)^{f-1}\beta_2^{g-1}(1-\beta_2)^{h-1}$$

Ahora considerando la función $L(\theta, \beta; d)$ y el planteamiento bayesiano se obtiene que la distribución a posteriori es la siguiente:

$$\pi(\theta_0, \beta_0, \beta_1, \beta_2|d) \propto \theta_0^{n_0(s)+a-1}(1-\theta_0)^{n_1(s)+b-1}\beta_0^{r_{0,2}+c-1}(1-\beta_0)^{r_{0,0}+d-1}\beta_1^{r_{1,2}+e-1}(1-\beta_1)^{r_{1,0}+f-1}$$
$$\beta_2^{r_{2,2}+g-1}(1-\beta_2)^{r_{2,0}+h-1}\left[\theta_0(1-\beta_0)^{n_0(s_0)}(1-\beta_1)^{n_1(s_0)} + (1-\theta_0)(1-\beta_1)^{n_0(s_0)}(1-\beta_2)^{n_1(s_0)}\right]^{n(\bar{s})}$$

El uso de la distribución a priori permite obtener una fórmula analítica del estimador Bayes y de la distribución marginal a posteriori. Como no se tiene conocimiento previo sobre la distribución a priori específica, Chow y Thompson [4] seleccionan tres de

esta familia y concluyen que el estimador Bayes no es sensitivo a cada una de ellas. Por lo tanto para la estimación de los paramétros se ha considerado $Beta(0.5, 0.5)$

Para calcular el estimador bayesiano de θ_0 se procede a evaluar la $E(\theta_0|d)$. Primero, se integra la distribución marginal

$$m(\theta_0, \beta_0, \beta_1, \beta_2)$$

es decir

$$M_1 = \int_0^1 \int_0^1 \int_0^1 \int_0^1 m(\theta_0, \beta_0, \beta_1, \beta_2) d\theta_0 d\beta_0 d\beta_1 d\beta_2$$

Luego se calcula esperanza para el parámetro θ_0 en la distribución marginal es decir,

$$M_2 = \int_0^1 \int_0^1 \int_0^1 \int_0^1 \theta_0 m(\theta_0, \beta_0, \beta_1, \beta_2) d\theta_0 d\beta_0 d\beta_1 d\beta_2$$

Entonces, el estimador de Bayes para θ_0 es la media de la distribución a posteriori, es decir

$$\hat{\theta}_0 = E(\theta_0|d) = \frac{M_2}{M_1},$$

Puesto que $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = B(\alpha, \beta)$, las expresiones para M_1 y M_2 son respectivamente:

$$M_1 = \sum_{i=0}^{n(\bar{s})} \binom{n(\bar{s})}{i} B(n_0(s)+a+i, n(\bar{s})+n_1(s)+b-i) B(r_{0,2}+c, i*n_0(s_0)+r_{0,0}+d)$$

$$B(r_{1,2}+e, i*n_1(s_0)+(n(\bar{s})-i)n_0(s_0)+r_{1,0}+f) B(r_{2,2}+g, (n(\bar{s})-i)n_1(s_0)+r_{2,0}+h)$$

y

$$M_2 = \sum_{i=0}^{n(\bar{s})} \binom{n(\bar{s})}{i} B(n_0(s)+a+1+i, n(\bar{s})+n_1(s)+b-i) B(r_{0,2}+c, i*n_0(s_0)+r_{0,0}+d)$$

$$B(r_{1,2}+e, i*n_1(s_0)+(n(\bar{s})-i)n_0(s_0)+r_{1,0}+f) B(r_{2,2}+g, (n(\bar{s})-i)n_1(s_0)+r_{2,0}+h)$$

Hasta ahora se ha mostrado todo el desarrollo para la proporción θ_0, pero es importante mencionar que la estimación de los demás parámetros sigue el mismo procedimiento. Como estamos considerando el caso simétrico, es decir un grafo no dirigido, sólo falta calcular β_2. Luego es necesario calcular la esperanza de β_2 en la distribución marginal. Entonces,

$$\hat{\beta}_2 = E(\beta_0|d) = \frac{M_3}{M_1},$$

donde,

$$M_3 = \sum_{i=0}^{n(\bar{s})} \binom{n(\bar{s})}{i} B(n_0(s)+a+1+i, n(\bar{s})+n_1(s)+b-i) B(r_{0,2}+c, i*n_0(s_0)+r_{0,0}+d)$$

$$B(r_{1,2}+e, i*n_1(s_0)+(n(\bar{s})-i)n_0(s_0)+r_{1,0}+f) B(r_{2,2}+g+1, (n(\bar{s})-i)n_1(s_0)+r_{2,0}+h)$$

Consideremos el problema de predecir por medio de la muestra de datos, alguna característica específica del grafo $z = f(y, x)$. Denotando los parámetros por ψ, la función general a posteriori para predecir en el grafo z es:

$$f(z|d) = \int f(z|d,\psi)\pi(\psi|d)d\psi \propto \int f(z,d|\psi)\pi(\psi)d\psi$$

donde la constante de proporcionalidad es $f(d)$. De nuevo $n_1(s)$ indica la cantidad de nodos cuyo valor de $y = 1$ y $n_1(\bar{s})$ denota los nodos con valor 1 entre los nodos que no están en la muestra, $n_1(s)$ es una cantidad observable mientras que $n_1(\bar{s})$ no es observable. Por lo tanto la proporción de nodos con valor 1 se estima con la expresión $z = \frac{n_1(s)+n_1(\bar{s})}{N}$. Este estimador implica que sólo $n_1(\bar{s})$ es visto como una variable aleatoria en lugar de fija, de ahí que se denomine "Bayes Predictor". Puesto que el anterior estimador tiende a producir un razonable punto de estimación, no

es tan útil como en la predicción de intervalos, debido a que los estimados de los parámetros con tratados como los verdaderos valores. La función para predecir el valor de la variable aleatoria $n_1(\bar{s})$ de la muestra observada es:

$$f(\theta,\lambda;d,n_1(\bar{s}))=p(s|y_s,x_{s_0U})\prod_i \theta_i^{n_i(s)+n_i(\bar{s})}\binom{n(\bar{s})}{n_1(\bar{s})}\prod_{ijkl}\lambda_{ijkl}^{m_{ijkl}(s_0,s_0)}\prod_{ijk}\lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0,s_1)}\prod_{ij}\lambda_{ij0\bullet}^{n_i(s_0)n_j(\bar{s})}$$

Y en el caso del modelo simétrico la función es:

$$f(\theta,\lambda;d,n_1(\bar{s}))=p(s|y_s,x_{s_0U})\prod_i \theta_i^{n_i(s)+n_i(\bar{s})}\binom{n(\bar{s})}{n_1(\bar{s})}\prod_{ij}\beta_{i+j}^{m_{ij00}(s_0)}(1-\beta_{i+j})^{m_{ij00}(s_0)+n_i(s_0)n_j(\bar{s})}$$

Desarrollando la productoria se tiene:

$$f(d,n_1(\bar{s})|\theta_0,\beta_0,\beta_1,\beta_2)=p(s|y_s,x_{s_0U})\binom{n(\bar{s})}{n_1(\bar{s})}$$
$$\theta_0^{n_0(s)+n_0(\bar{s})}(1-\theta_0)^{n_1(s)+n_1(\bar{s})}\beta_0^{r_{02}(s_0,s)}(1-\beta_0)^{r_{00}(s_0,s)+n_0(s_0)n_0(\bar{s})}$$
$$\beta_1^{r_{12}(s_0,s)}(1-\beta_1)^{r_{10}(s_0,s)+n_0(s_0)n_1(\bar{s})+n_1(s_0)n_0(\bar{s})}\beta_2^{r_{22}(s_0,s)}(1-\beta_2)^{r_{20}(s_0,s)+n_1(s_0)n_1(\bar{s})}$$

Desarrollando la integral mediante las distribuciones independientes a priori de los parámetros, la distribución posterior es:

$$f(n_1(\bar{s})|d)\propto\binom{n(\bar{s})}{n_1(\bar{s})}*B[n_0(s)+n_0(\bar{s})+a,n_1(s)+n_1(\bar{s})+b]*B[r_{02}+c,r_{00}+n_0(s_0)n_0(\bar{s})+d]$$
$$B[r_{12}+e,r_{10}+n_0(s_0)n_1(\bar{s})+n_1(s_0)n_0(\bar{s})+f]*B[r_{22}+g,r_{20}+n_1(s_0)n_1(\bar{s})+h]$$

Ahora es necesario calcular la esperanza de $n_1(\bar{s})$ entonces

$$E(n_1(\bar{s})|d)=\sum_{n_1(\bar{s})}^{n(\bar{s})} n_1(\bar{s})f(n_1(\bar{s}))|d)$$

como

$$i * \binom{n}{i} = n * \binom{n-1}{i-1}$$

Por lo tanto

$$E[n_1(\bar{s})|d] \propto n(\bar{s}) \sum_{i=1}^{n(\bar{s})} \binom{n(\bar{s})-1}{i-1}$$

$$B[n_0(s) + n(\bar{s}) + a - i, n_1(s) + b + i] * B[r_{02} + c, r_{00} + n_0(s_0)(n(\bar{s}) - i) + d]$$

$$B[r_{12}+e, r_{10}+n_0(s_0)*i+n_1(s_0)(n(\bar{s})-1)+f]*B[r_{22}+g, r_{20}+n_1(s_0)*i+h] = M_4$$

Así que la $E[n_1(\bar{s})|d] = M_4/M_1$, por lo tanto el estimador Bayes de la proporción de nodos positivo es

$$\hat{z} = E(z|d) = E\left[\frac{n_1(s) + n_1(\bar{s})}{N}|d\right] = \frac{n_1(s) + (M_4/M_1)}{N}$$

4.3 Una demostración numérica con datos reales

Esta demostración se realiza con el fin verificar y comprobar la estructura correcta del algoritmo implementado en **R** para obtener los estimados como lo plantea la teoría anterior expuesta, y luego hacer las variaciones relacionadas al objetivo propuesto en este trabajo, considerando la estratificación de la población.

El conjunto de datos utilizado se origina de un estudio realizado en Colorado Springs sobre el contagio de HIV/AIDS en miembros heterosexuales [29]. Los datos son recopilados desde el año 1981 hasta 1992, de centros médicos privados, públicos, militares y en centros de donación de sangre, además de prostitutas y drogadictos. A cada unos de los individuos se les lleva un registro con sus contactos sexuales al igual que sus compañeros de consumo de drogas. Entre los datos de Colorado Springs se conoce que el número promedio de prostitutas por cada año está entre 100 y 120, y que el promedio de drogadictos es aproximadamente de 1500 según el información nacional. Desde 1985 se ha tenido seguimiento tanto de prostitutas como de sus

clientes y de los que no lo son. Exactamente se tienen 595 individuos y alrededor de más de 2000 relaciones componen la red entre sexuales y de drogadicción.

El archivo original de los datos contiene información en diferentes aspectos (socio-económico, geográfico, personal) y es necesario identificar, extraer y ordenar la lista para realizar la implementación en **R** de tal forma que sea eficiente. La lista de los contactos está en el archivo red595.txt y la lista de los individuos en riesgo de trasmitir y ser contagiados está en el archivo indsexwork.txt. El algoritmo está compuesto por la parte a continuación y por el algoritmo 1 implementados en **R**:

```
library(sna)
library(network)
gr=read.table("C:/Documents and Settings/maria/Desktop/Willi/+
Datos/red595.txt",head=T)
gr=as.matrix.network(network(gr),matrix.type="adjacency")
gr=symmetrize(gr)
u=read.table("C:/Documents and Settings/maria/Desktop/Willi/Datos/+
indsexwork.txt",head=T)
u=u$ind
x=c(1:595)
nd=as.matrix.network(network(gr),matrix.type="edgelist")
dyad.census(gr)
nr=100
propu=length(u)/length(x)
```

Después de ejecutar el programa en **R** se tiene los siguientes resultados que son muy satisfactorios para hacer inferencia. El número de iteraciones fue de 100, la proporción poblacional es 0.2235 y la probabilidad de que exista un arco entre dos vértices positivos es de 0.03201; los estimados son los siguientes:

Table 4–1: Estimación en Colorado Springs

estimador	θ_1	θ_{1pr}	**prop**	β_2
media	0.222578	0.222112	0.015042	0.035133
mse	0.002076	0.002084	0.043488	0.000360

Se puede apreciar que los dos tipos de estimadores por "Rastreo de vínculos" (θ_1=0.222578,θ_{1pr}=0.222112) tiene mejor aproximación al parámetro en comparación

con el estimador usual (prop=0.015042) además que sus respectivos mse son bajos. Por otra parte el estimador para β_2 tiene un mse bajo, lo cual implica que la diferencia del parámetro real y estimador es pequeña.

CAPÍTULO 5
RESULTADOS DEL MÉTODO ESTRATIFICADO-BOLA DE NIEVE

Una forma que podría disminuir el error de un estimador, es dividir la población en grupos hetereogeneos con características muy similares en cada uno de ellos, es decir en estratos. Teniendo en cuenta un conocimiento previo de los factores que pueden generar éstas divisiones, se pueden obtener estimadores parciales de la población e identificar cuales tiene mayor influencia en el estimador total.

Con el objetivo de encontrar un estimador que pueda disminuir variabilidad y optimizar los medios para obtener una muestra representativa, se propone dividir la población en grupos disyuntos llamados estratos y aplicar el método "bola de nieve" expuesto anteriormente. Primero es necesario exponer las definiciones y fundamentos de este tipo de muestreo.

El muestreo estratificado se obtiene de separar la población en grupos disyuntos, en donde se selecciona una muestra proporcional al tamaño del estrato por muestreo simple. La razón principal del empleo de esta técnica es disminuir el error de estimación aprovechando la similitud de los grupos e identificar qué grupos o estratos específicos influyen más en la estimación del parámetro en general. Algunas notaciones requeridas en este tipo de muestreo son las siguientes:

L = Número de estratos

N_i=Número de individuos por estrato

N=Número de individuos en la población.

$N = N_1 + N_2 + ... + N_L$

Sea $\bar{y}_i$, n_i, μ_i, τ_i, la media muestral, el tamaño de la muestra, la media poblacional y la poblacion total respectivamente por muestreo aleatorio simple (MAS) para el estrato i. La población total es $\tau = \sum_{i=1}^{L} \tau_i$. Como aplicamos (MAS) en cada estrato se sabe que $\bar{y}_i$ es un estimador insesgado para μ_i y por lo tanto $N_i\bar{y}_i$ es un estimador insesgado para la población del estrato i. Luego la fórmulas son:
Estimador de la población media μ.

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i\bar{y}_i$$

Estimador de la varianza de $\bar{y}_{st}$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$$

donde, $s_i^2 = \frac{1}{n_i-1}\sum_{i=1}^{N_i}(y_i - \bar{y}_i)^2$.
El estimador de la población total τ es:

$$N\bar{y}_{st} = \sum_{i=1}^{L} N_i\bar{y}_i$$

y la varianza de $N\bar{y}_{st}$ es

$$\hat{V}(N\bar{y}_{st}) = N^2\hat{V}(\bar{y}_{st}) = \sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$$

Para estimar la proporción de la población, se determina inicialmente la proporción $\hat{p}_i$ de cada uno de los estratos. Este estimador $\hat{p}_i$ es un estimador insesgado y por lo tanto $N_i\hat{p}_i$ es un estimador insesgado de la proporción de la población. Entonces el estimador de la población es:

$$\hat{p}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i\hat{p}_i$$

y la varianza de $\hat{p}_{st}$ es:

$$\hat{V}(\hat{p}_{st}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \left(\frac{\hat{p}_i \hat{q}_i}{n_i - 1} \right)$$

Los resultados presentados en este capítulo son obtenidos por medio de la base teórica original presentada por Thompson y Frank [10] junto con las variaciones que se refieren a la división de la población en estratos, que es el objetivo principal de este trabajo.

Se prueban cuatro clases de algoritmos de las siguientes características:

- El algoritmo 1 refleja todo el planteamiento de la teoría enunciada por Thompson y Frank. Consiste en seleccionar una muestra de forma aleatoria simple, obtener el número de contactos y los estadísticos respectivos. Es importante tener en cuenta que de una u otra forma los individuos seleccionados son identificados para evitar la multiplicidad. La forma de ser identificados la define el investigador por medio de alguna metodología.
- En el algoritmo 2 se combina la teoría de muestreo por estratos con el algoritmo 1. La población se divide en los estratos respectivos. En cada estrato se toma una muestra aleatoria proporcional al tamaño del estrato. Para cada S_0 se obtiene S_1. Luego se obtiene el conjunto total de todos los S_1, para compararlos con la muestra total inicial y así extraer de la cuenta a los que posiblemente estén en S_0. A partir de esta onda final se obtienen los estadísticos y por lo tanto el estimador de la proporción. El fin de esta variación es encontrar el mayor número de contactos teniendo un conocimiento previo, y obtener mayor eficiencia en la cobertura. Aquí los contactos son mencionados de toda la población, dentro o fuera del estrato.
- El algoritmo 3 de nuevo divide la población en estratos, pero sólo los contactos se consideran dentro de un mismo estrato. Todos los contactos S_1 se reunen, y no es necesario identificarlos con la muestra inicial pues cada estrato es disjunto. En cada estrato se calculan todos los estadísticos excepto el estimador. Se suman

los todos diferentes estadísticos de los respectivos estratos y luego se calcula el estimador final. La característica de este estimador es que sólo identificamos los contactos en los estratos y no en toda la población.

- El algoritmo 4 es básicamente en esencia la teoría y desarrollo del muestreo estratificado. Se tienen los estratos y se cuantifica la cantidad de contactos sin importar la identificación de los individuos seleccionados, se calculan los estadísticos y se obtiene el estimador respectivo a cada estrato. Luego, se obtiene el estimador de la proporción de toda la población, como se mencionó anteriormente, para estimar la proporción en muestreo estratificado. Los estimadores de cada estrato son independientes entre sí. Las múltiples relaciones no son consideradas y sólo nos interesa la cantidad de contactos referidos.

5.1 Simulaciones y resultados

A continuación se muestran los resultados obtenidos en cien simulaciones realizadas a los cuatro algoritmos planteados, en dos tipos de conjuntos, en los cuales se ha modificado el tamaño de la muestra inicial y la proporción de la subpoblación de interés. En cada uno de los conjuntos y algoritmos se han calculado los dos tipos de estimadores de Bayes, θ_1 (Estimador bayes de individuos con la característica), θ_{1pr} (Estimador bayes predictor, con individuos con la característica), y el estimador usual de proporción, con sus respectivos **mse** (error mínimo cuadrático), para apreciar la aproximación del parámetro estimado al parámetro real.

El primer conjunto, DATA1, es un grafo no dirigido con 500 nodos de los cuales 154 de ellos son aislados, además la cantidad de relaciones mutuas existentes entre ellos es de 612 de 124750 posibles y por consiguiente una densidad de 0.004905812.

La forma de generar este grafo en R es la siguiente forma:

```
gr=rgraph(350,1,0.01,mode="graph")
gr=add.isolates(gr,150)
gr=rmperm(gr)
```

El segundo conjunto, DATA2, es un grafo también no dirigido compuesto de 1000 nodos, de los cuales 100 son aislados. El número de relaciones mutuas existentes entre ellos es de 3980 de 499500 posibles. Los comandos para generar el este grafo son los siguientes:

```
gr=rgraph(900,1,0.01,mode="graph")
isolates(gr)=100
gr=rmperm(gr)
gden(gr)=0.007967968
```

5.2 Resultados obtenidos DATA 1.

Table 5–1: Proporción 0.1, grafo de 500 nodos

Promedio de los estimadores de la proporción en la 100 muestras por medio de los Algoritmo 1 (S_0 Población), Algoritmo 2 (S_0 en cada estrato), Algoritmo 3 (S_0, S_1 en cada estrato) y Algoritmo 4 (Estimador parcial de cada estrato con S_0, S_1) con diferente S_0 y con proporción=0.1, en un grafo de 500 nodos

S_0	**Estimador**	**Algoritmo 1**	**Algoritmo 2**	**Algoritmo 3**	**Algoritmo 4**
50	θ_1	0.091453	0.093346	0.090989	0.092119
	θ_{1pr}	0.090636	0.092533	0.090169	0.091303
	prop	0.010340	0.009860	0.009340	0.002614
40	θ_1	0.090585	0.093162	0.099163	0.093889
	θ_{1pr}	0.089766	0.092349	0.098360	0.093076
	prop	0.007940	0.008440	0.007940	0.002253
25	θ_1	0.090005	0.089416	0.093885	0.100142
	θ_{1pr}	0.089185	0.088595	0.093059	0.099343
	prop	0.005040	0.004580	0.005100	0.001432

Table 5–2: Proporción 0.05, grafo de 500 nodos

Promedio de los estimadores de la proporción en la 100 muestras por medio de los Algoritmo 1 (S_0 Población), Algoritmo 2 (S_0 en cada estrato), Algoritmo 3 (S_0, S_1 en cada estrato) y Algoritmo 4 (Estimador parcial de cada estrato con S_0, S_1) con diferente S_0 y con proporción=0.05, en un grafo de 500 nodos

S_0	**Estimador**	**Algoritmo 1**	**Algoritmo 2**	**Algoritmo 3**	**Algoritmo 4**
50	θ_1	0.05175	0.050384	0.047063	0.056151
	θ_{1pr}	0.050854	0.049485	0.046110	0.055263
	prop	0.005120	0.005040	0.004920	0.001540
40	θ_1	0.049601	0.051814	0.046471	0.059810
	θ_{1pr}	0.048700	0.050917	0.045496	0.058929
	prop	0.003740	0.004060	0.004040	0.001236
25	θ_1	0.050623	0.047808	0.049149	0.064856
	θ_{1pr}	0.049724	0.046904	0.048172	0.063986
	prop	0.002600	0.002360	0.002400	0.000742

Table 5–3: Proporción 0.02, grafo de 500 nodos

Promedio de los estimadores de la proporción en la 100 muestras por medio de los Algoritmo 1 (S_0 Población), Algoritmo 2 (S_0 en cada estrato), Algoritmo 3 (S_0, S_1 en cada estrato) y Algoritmo 4 (Estimador parcial de cada estrato con S_0, S_1) con diferente S_0 y con proporción=0.02, en un grafo de 500 nodos

S_0	**Estimador**	**Algoritmo 1**	**Algoritmo 2**	**Algoritmo 3**	**Algoritmo 4**
50	θ_1	0.020279	0.020140	0.019944	0.029239
	θ_{1pr}	0.019319	0.019180	0.018807	0.028297
	prop	0.001920	0.001880	0.001740	0.000592
40	θ_1	0.021674	0.021746	0.023394	0.032810
	θ_{1pr}	0.020717	0.020789	0.022281	0.031875
	prop	0.001520	0.001400	0.001800	0.000442
25	θ_1	0.024538	0.023427	0.026172	0.038256
	θ_{1pr}	0.023588	0.022474	0.025063	0.037332
	prop	0.000820	0.001100	0.000980	0.000210

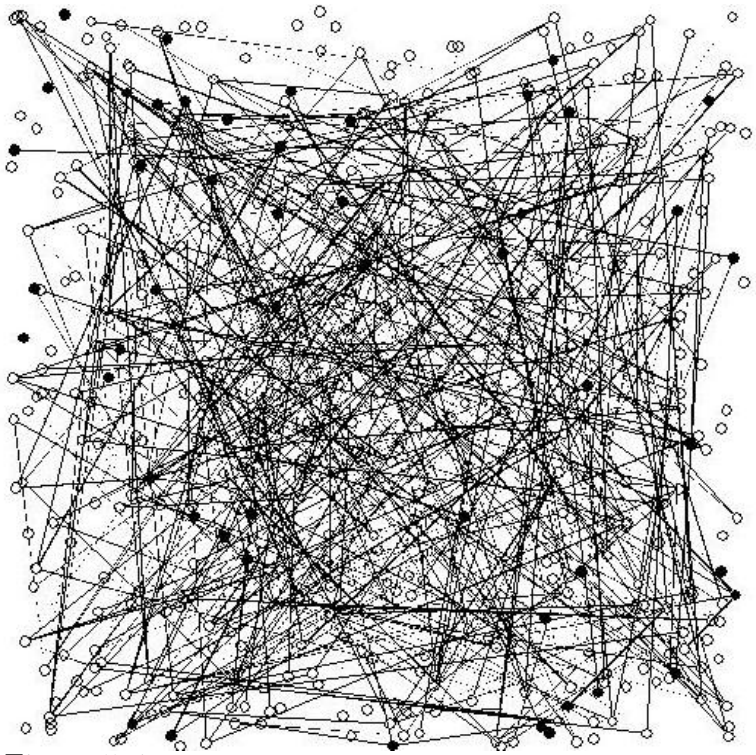

Figure 5–1: Grafo de DATA1 con 500 nodos en modo random

La información ofrecida en la figura 5-1 es de carácter descriptivo. Este grafo representa la población de 500 vértices con sus respectivas relaciones entre ellos. Es un grafo no dirigido. Los vértices de color negro representan la subpoblación y el resto de ellos está de color blanco o demarcados. La distribución de los vértices obedece a las condiciones que ofrece el modo "random" en **R** y por lo tanto los vértices aislados no son muy facilmente aprecidados. Es importante notar que la estructura tiende a ser compleja y por ende el análisis de dichas interacciones.

La figura 5-2 representa la misma población con la diferencia de que los vértices aislados se pueden apreciar facilmente.

La figura 5-3 representa una muestra de 50 miembros con sus contactos, identificados por un círculo y un triángulo respectivamente. Los individuos que tienen la característica están de color negro, a diferencia de los demás que son blancos. En ella se puede apreciar con más detalle sus respectivas conexiones.

La figura 5-4 sólo muestra la poco información que podemos obtener a partir la

Table 5–4: Error cuadrático medio de los estimadores en DATA1

Error cuadrático medio (**mse**) de cada uno de los estimadores de la proporción en la 100 muestras por medio de los Algoritmo 1 (S_0 Población), Algoritmo 2 (S_0 en cada estrato), Algoritmo 3 (S_0, S_1 en cada estrato) y Algoritmo 4 (Estimador parcial de cada estrato con S_0, S_1) con diferente S_0 y proporciones, en un grafo de 500 nodos

Prop	S_0	**Estimador**	**Algoritmo 1**	**Algoritmo 2**	**Algoritmo 3**	**Algoritmo 4**
		θ_1	0.000441	0.000533	0.000823	0.000465
	50	θ_{1pr}	0.000458	0.000546	0.000842	0.000480
		prop	0.008062	0.008140	0.008232	0.009485
		θ_1	0.000514	0.000403	0.000850	0.000521
0.1	40	θ_{1pr}	(0.000532	0.000417	0.000855	0.000534
		prop	0.008489	0.008395	0.008487	0.009556
		θ_1	0.000815	0.000746	0.001610	0.000798
	25	θ_{1pr}	0.000835	0.000766	0.001628	0.000801
		prop	0.009024	0.009110	0.009014	0.009716
		θ_1	0.000238	0.000167	0.000499	0.000199
	50	θ_{1pr}	0.000236	0.000168	0.000510	0.000190
		prop	0.002022	0.002030	0.002042	0.002349
		θ_1	0.000268	0.000191	0.000676	0.000355
0.05	40	θ_{1pr}	0.000270	0.000189	0.000690	0.000339
		prop	0.002147	0.002118	0.002122	0.002379
		θ_1	0.000426	0.000373	0.000847	0.000570
	25	θ_{1pr}	0.000427	0.000379	0.000856	0.000546
		prop	0.002251	0.002274	0.002270	0.002427
		θ_1	0.000075	0.000076	0.000114	0.000171
	50	θ_{1pr}	0.000075	0.000077	0.000118	0.000154
		prop	0.000331	0.000332	0.000337	0.000377
		θ_1	0.000089	0.000098	0.000188	0.000285
0.02	40	θ_{1pr}	0.000087	0.000096	0.000186	0.000263
		prop	0.000344	0.000349	0.000334	0.000383
		θ_1	0.000187	0.000166	0.000271	0.000472
	25	θ_{1pr}	0.000180	0.000161	0.000262	0.000439
		prop	0.000370	0.000360	0.000363	0.000392

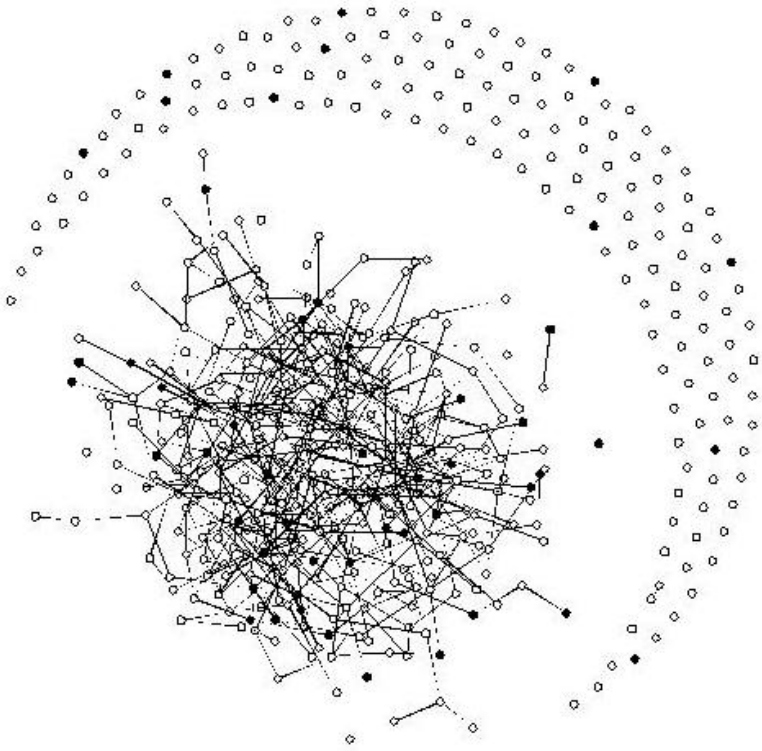

Figure 5–2: Grafo de DATA1 con 500 nodos en modo estándar

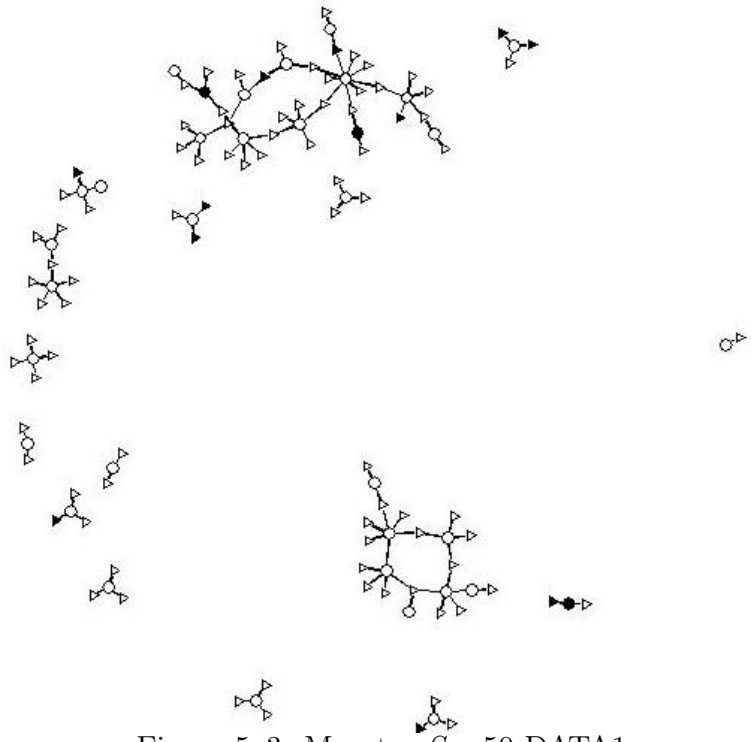

Figure 5–3: Muestra S_0=50 DATA1
Proporción=0.1, los nodos redondos indican S_0, los nodos negros los individuos de interés y los triángulos S_1

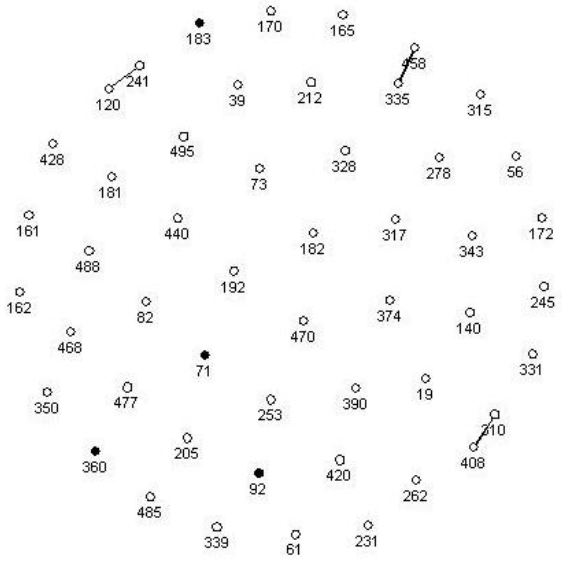

Figure 5–4: Muestra $S_0 \times S_0$ Proporción=0.1

muestra inicial sin aplicar "bola de nieve".

La figura 5-5 es la gráfica de la red de contactos de una muestra inicial de $S_0 = 25$, al igual que la figura 5-3 la muestra y los contactos se representan por círculos y triángulos respectivamente.

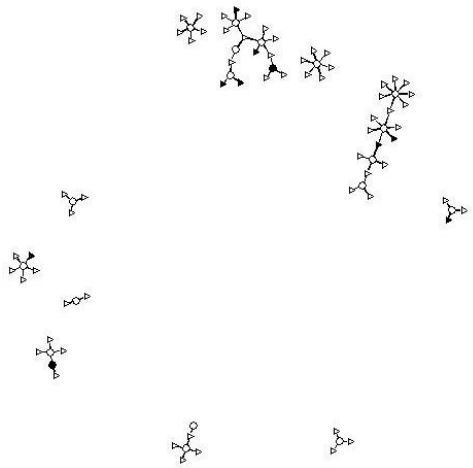

Figure 5–5: Muestra S_0=25 Proporción=0.1

5.3 Resultados obtenidos DATA 2.

Table 5–5: Proporción 0.1, grafo de 1000 nodos

Promedio de los estimadores de la proporción en la 100 muestras por medio de los Algoritmo 1 (S_0 Población), Algoritmo 2 (S_0 en cada estrato), Algoritmo 3 (S_0, S_1 en cada estrato) y Algoritmo 4 (Estimador parcial de cada estrato con S_0, S_1) con diferente S_0 y con proporción=0.1, en un grafo de 1000 nodos

S_0	**Estimador**	**Algoritmo 1**	**Algoritmo 2**	**Algoritmo 3**	**Algoritmo 4**
50	θ_1	0.089318	0.088460	0.097349	0.088493
	θ_{1pr}	0.088907	0.088049	0.096947	0.088082
	prop	0.005130	0.005100	0.004370	0.001498
40	θ_1	0.088617	0.089295	0.102932	0.087486
	θ_{1pr}	0.088206	0.088884	0.102535	0.087073
	prop	0.004090	0.004180	0.004200	0.001148
25	θ_1	0.088647	0.088000	0.097119	0.093170
	θ_{1pr}	0.088235	0.087588	0.096716	0.092763
	prop	0.002110	0.002640	0.002290	0.000733

Table 5–6: Proporción 0.05, grafo de 1000 nodos

Promedio de los estimadores de la proporción en la 100 muestras por medio de los Algoritmo 1 (S_0 Población), Algoritmo 2 (S_0 en cada estrato), Algoritmo 3 (S_0, S_1 en cada estrato) y Algoritmo 4 (Estimador parcial de cada estrato con S_0, S_1) con diferente S_0 y con proporción=0.05, en un grafo de 1000 nodos

S_0	**Estimador**	**Algoritmo 1**	**Algoritmo 2**	**Algoritmo 3**	**Algoritmo 4**
50	θ_1	0.047138	0.046097	0.050017	0.047917
	θ_{1pr}	0.046685	0.045644	0.049566	0.047465
	prop	0.002500	0.002500	0.002400	0.000638
40	θ_1	0.045355	0.047661	0.049160	0.047918
	θ_{1pr}	0.04490	0.047209	0.048707	0.047466
	prop	0.002020	0.002010	0.002160	0.000536
25	θ_1	0.048639	0.047052	0.052746	0.051046
	θ_{1pr}	0.048188	0.046599	0.052296	0.050597
	prop	0.001160	0.001300	0.001370	0.000353

Table 5–7: Proporción 0.025, grafo de 1000 nodos

Promedio de los estimadores de la proporción en la 100 muestras por medio de los Algoritmo 1 (S_0 Población), Algoritmo 2 (S_0 en cada estrato), Algoritmo 3 (S_0, S_1 en cada estrato) y Algoritmo 4 (Estimador parcial de cada estrato con S_0, S_1) con diferente S_0 y con proporción=0.025, en un grafo de 1000 nodos

S_0	**Estimador**	**Algoritmo 1**	**Algoritmo 2**	**Algoritmo 3**	**Algoritmo 4**
50	θ_1	0.027787	0.027805	0.028607	0.032298
	θ_{1pr}	0.027314	0.027333	0.028125	0.031830
	prop	0.001450	0.000990	0.001270	0.000341
40	θ_1	0.029463	0.029385	0.027740	0.032702
	θ_{1pr}	0.028993	0.028914	0.027251	0.032235
	prop	0.001000	0.000990	0.000970	0.000263
25	θ_1	0.028676	0.029705	0.030563	0.035653
	θ_{1pr}	0.028205	0.029235	0.030068	0.035189
	prop	0.000570	0.000620	0.000540	0.000154

Table 5–8: Proporción 0.01, grafo de 1000 nodos

Promedio de los estimadores de la proporción en la 100 muestras por medio de los Algoritmo 1 (S_0 Población), Algoritmo 2 (S_0 en cada estrato), Algoritmo 3 (S_0, S_1 en cada estrato) y Algoritmo 4 (Estimador parcial de cada estrato con S_0, S_1) con diferente S_0 y con proporción=0.01, en un grafo de 1000 nodos

S_0	**Estimador**	**Algoritmo 1**	**Algoritmo 2**	**Algoritmo 3**	**Algoritmo 4**
50	θ_1	0.011649	0.011571	0.011903	0.014593
	θ_{1pr}	0.011161	0.011083	0.011344	0.014107
	prop	0.000490	0.000540	0.000470	0.000100
40	θ_1	0.011147	0.011917	0.012892	0.016576
	θ_{1pr}	0.010659	0.011429	0.012330	0.016093
	prop	0.000290	0.000420	0.000400	0.000114
25	θ_1	0.011061	0.011618	0.016472	0.019055
	θ_{1pr}	0.010572	0.011130	0.015920	0.018574
	prop	0.000260	0.000270	0.000260	0.000051

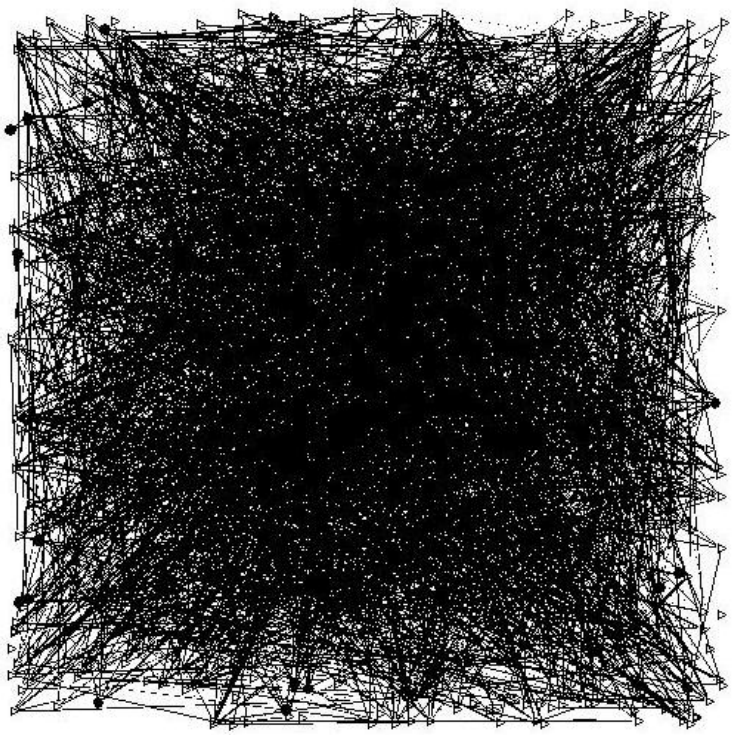

Figure 5–6: Grafo de DATA2 con 1000 nodos en modo random

Table 5–9: Error cuadrático medio en los estimadores de DATA2

Error cuadrático medio (**mse**) de cada uno de los estimadores de la proporción en la 100 muestras por medio de los Algoritmo 1 (S_0 Población), Algoritmo 2 (S_0 en cada estrato), Algoritmo 3 (S_0, S_1 en cada estrato) y Algoritmo 4 (Estimador parcial de cada estrato con S_0, S_1) con diferente S_0 y proporciones, en un grafo de 1000 nodos

Prop	S_0	**Estimador**	**Algoritmo 1**	**Algoritmo 2**	**Algoritmo 3**	**Algoritmo 4**
		θ_1	0.000274	0.000243	0.000483	0.000253
	50	θ_{1pr}	0.000283	0.00025	0.000486	0.000263
		prop	0.009005	0.009011	0.009149	0.009703
		θ_1	0.000291	0.000304	0.000586	0.000323
0.1	40	θ_{1pr}	0.000301	0.000314	0.000585	0.000334
		prop	0.009201	0.009185	0.00918	0.009772
		θ_1	0.000473	0.00040	0.000808	0.000333
	25	θ_{1pr}	0.000483	0.000414	0.000813	0.000339
		prop	0.009584	0.009481	0.009549	0.009854
		θ_1	0.000099	0.000090	0.000249	0.000082
	50	θ_{1pr}	0.000102	0.000094	0.000249	0.000085
		prop	0.002259	0.002258	0.002268	0.002437
		θ_1	0.000116	0.000098	0.000286	0.000104
0.05	40	θ_{1pr}	0.000121	0.000101	0.000288	0.000107
		prop	0.002304	0.002305	0.002290	0.002447
		θ_1	0.000183	0.000167	0.000435	0.000148
	25	θ_{1pr}	0.000185	0.000170	0.000434	0.000148
		prop	0.002386	0.002373	0.00236	0.002465
		θ_1	0.000050	0.000050	0.000127	0.000109
	50	θ_{1pr}	0.000048	0.000047	0.000124	0.000103
		prop	0.000556	0.000577	0.000564	0.000608
		θ_1	0.000066	0.000078	0.000151	0.000115
0.025	40	θ_{1pr}	0.000062	0.000074	0.000149	0.000108
		prop	0.000576	0.000577	0.000578	0.000612
		θ_1	0.000106	0.000104	0.000294	0.000237
	25	θ_{1pr}	0.000103	0.000100	0.000291	0.000228
		prop	0.000597	0.000595	0.000599	0.000617
		θ_1	0.000019	0.000019	0.000040	0.000039
	50	θ_{1pr}	0.000018	0.000018	0.000038	0.000035
		prop	0.000091	0.000090	0.000091	0.000098
		θ_1	0.000019	0.000029	0.000074	0.000070
0.001	40	θ_{1pr}	0.000018	0.000027	0.000072	0.000064
		prop	0.000095	0.000092	0.000093	0.000098
		θ_1	0.000041	0.000040	0.000156	0.000113
	25	θ_{1pr}	0.000040	0.000039	0.000151	0.000104
		prop	0.000095	0.000095	0.000095	0.000099

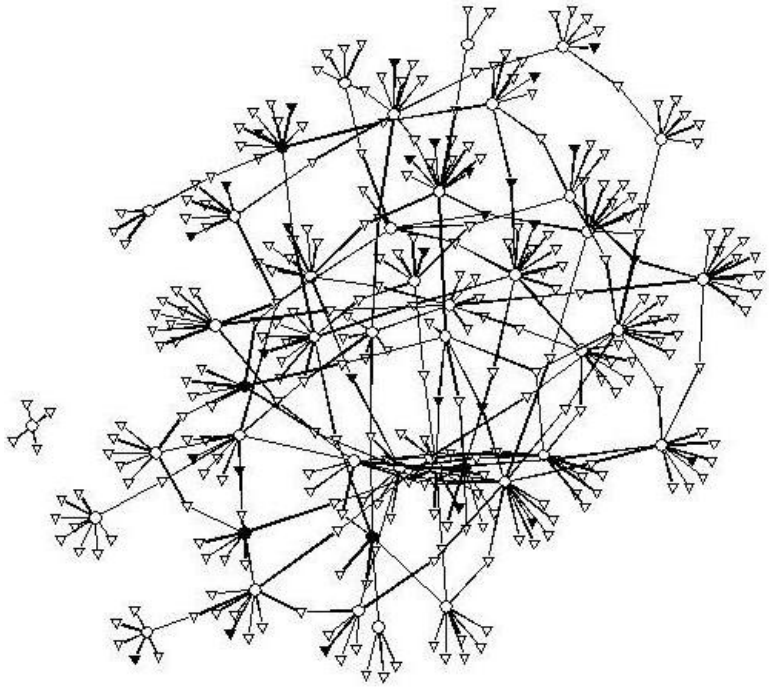

Figure 5–7: Muestra S_0=50 DATA2
Proporción=0.1, los nodos redondos indican S_0, los nodos negros los individuos de interés y los triángulos S_1

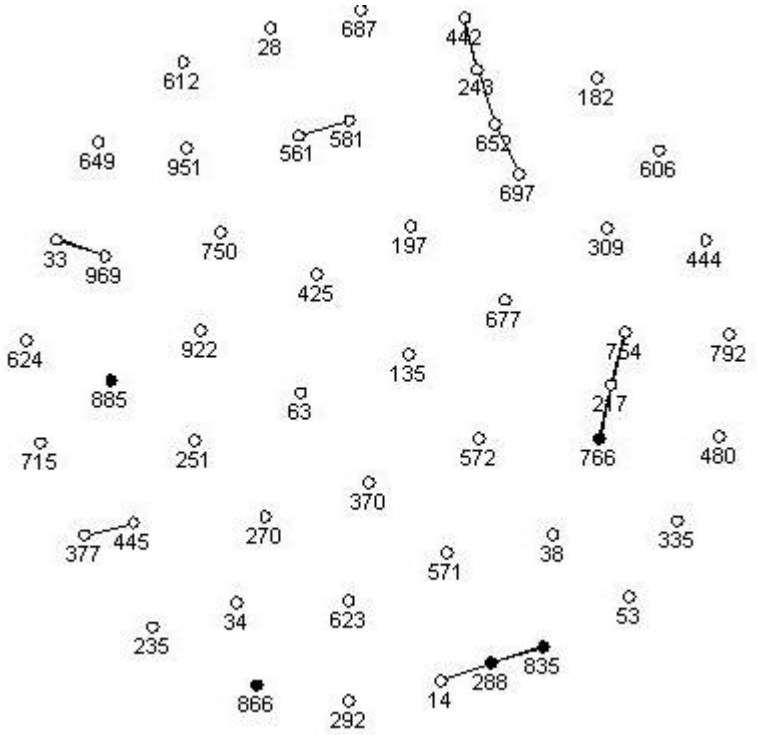

Figure 5–8: Muestra $S_0 \times S_0$ Proportion=0.1

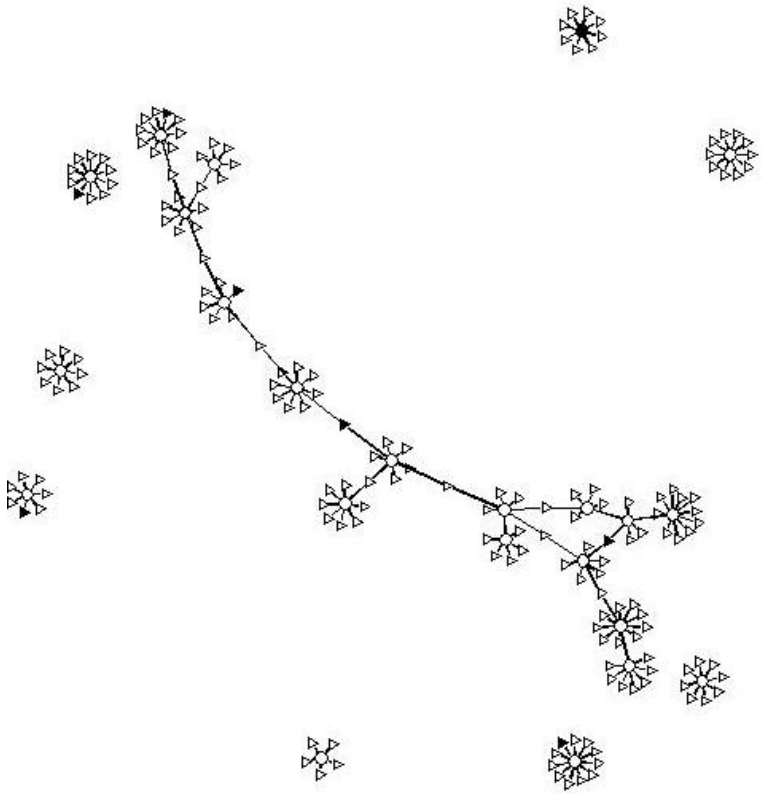

Figure 5–9: Muestra S_0=50 Proportion=0.1
Proporción=0.1, los nodos redondos indican S_0, los nodos negros los individuos de interés y los triángulos S_1

La figura 5-6 representa un grafo de la DATA2, compuesto por 1000 nodos. Los vértices circulares de color negro representan los individuos con la característica, los demás están representados por un triángulo blanco.

La figura 5-7 muestra los contactos que genera una muestra de 50 vértices. La muestra inicial S_0 está representado por los círculos, los contactos por un triángulo y los que tienen la característica están de color oscuro.

La figura 5-8 representa la muestra inicial sin los contactos, es decir, la representacion de la proporcion de la subpoblacion.

Si comparamos la figura 5-8 con la figura 5-7 se puede apreciar que la proporción de individuos con la característica es menor con respecto a los vértices de color negro expuestos en la figura 5-7, lo que indica que el muestreo convencional no es apropiado.

La figura 5-9 es la red de contactos de una muestra de 25 vértices con sus respectivos contactos.

En las tablas 5-4 y 5-9, se puede observar que el estimador de proporción usual ofrece un **mse** alto en comparación con los **mse** de los estimadores del muestreo "Rastreo por vínculos", lo cual reafirma que los muestreos convencionales no son convenientes para este tipo de poblaciones.

CAPÍTULO 6
CONCLUSIONES Y TRABAJOS FUTUROS

6.1 Conclusiones

Después de llevar a cabo nuestro objetivo de obtener estimadores de la proporción de una población oculta, considerando muestreo estratificado con "Rastreo por vínculos", se concluyó lo siguiente:

Los estimados obtenidos por la combinación de muestreo estratificado con "Rastreo por vínculos", ofrece generalmente mejores estimados, basándonos en los mse calculados.

Con base en el error cuadrático medio (mse), se demostró que los métodos de muestreo convencionales no son adecuados en este tipo de poblaciones, debido a que sus valores obtenidos son muy altos en comparación con obtenidos en el muestreo adaptativo propuesto.

La combinación de muestreo estratificado con "Rastreo por vínculos", permite obtener estimadores parciales para cada estrato e identificar que cual de ellos influyen más en el parámetro total de la población.

Las diferentes condiciones en cada uno de los algoritmos permite considerar los posibles tipos de vínculos existentes en los estratos establecidos y así obtener un estimador con más precisión.

De acuerdo a las pruebas simuladas realizadas, se puede apreciar que los tamaños de la muestra inicial considerados no afectan de manera significativa el parámetro estimado.

6.2 Contribuciones

Las principales contribuciones después de la realización de este trabajo y las diversas pruebas efectuadas fueron:

- Exponer en forma clara y detallada la estimación del tamaño de una población oculta mediante muestro estratificado con "Rastreo por vínculos".
- Implementar y derrosallar en **R** los diferentes algoritmos en la estimación de parámetros en poblaciones ocultas.
- Promover la realización de estudios posteriores en poblaciones ocultas con datos e información obtenida en investigaciones.

6.3 Trabajos futuros

Como trabajos futuros se plantea trabajar con un modelo no simétrico, en el cual la existencia de un arco mutuo entre nodos no necesariamente se tiene; por lo tanto, la representación correspondería a un grafo dirigido, que en términos de probabilidades significa que $\lambda_{ijkl} \neq 0$ para $k \neq l$. En consecuencia, la presentación de la distribución final del modelo y los estadísticos para obtener los estimados cambiarían, por lo que el algoritmo en **R** sería diferente por la presencia de una matriz que en este caso sería no simétrica.

Quedó pendiente la evaluación de la probabilidad de arco mutuo β_2 utilizando poblaciones simuladas y la aplicación del método de muestreo estratificado y de "Rastreo por vínculos" a conjuntos de datos reales.

REFERENCIAS BIBLIOGRÁFICAS

REFERENCIAS BIBLIOGRÁFICAS

[1] Ove. Frank and T.A.B. Snijders. Estimating the Size of Hidden Populations Using Snowball Sampling. *Journal of Official Statistics*, 10:53–67, 1994.

[2] Martha Romero; Eva Maria Rodrguez; Ana Durand-Smith; Rosa Maria Aguilera. Veinticinco años de investigación cualitativa en salud mental y adicciones con poblaciones ocultas. *Salud Mental*, 26:76–83, Diciembre 2003.

[3] Linda M. Collins Steven K. Thompson. Adaptive Sampling in Research on Risk-Related Behaviors. *Drug and Alcohol Dependence*, 68:S57–S67, 2002.

[4] Thompson S.K. Chow, M. Estimation with Link-Tracing Sampling Designs, a Bayesian Approach. *Survey Methodology*, 29:197–205, 2003.

[5] Rothenberg Richard B. Commentary: Sampling in Social Networks. *Connectiones, INSNA*, 18:104–110, 1995.

[6] W. Wayne Wiebel. Identifying and Gaining Access to Hidden Populations. *The Collection and Interpretation of Data from Hidden Populations*, 98:4–11, 1990.

[7] Elizabeth Y; Lambert; W.Wayne Wiebel. *NIDA Research Monograph,The Collection and Interpretation of Data from Hidden Populations*, 98:1–4, 1990.

[8] Heckarthon D. Responden-Driven Sampling:a New Approach to the Study of Hidden Population. *Social Problems*, 44:2, 1997.

[9] Czaja R; Trunzo D; Royston P. Response Efects in a Network Sampling. *Sociological Methods y Research*, 20, No 3:340–366, 1992.

[10] Thompson S.K; Frank Ove. Model-Based Estimation with Link-Tracing Sampling Designs. *Survey Methodology*, 26:87–98, 2000.

[11] Laumann E.O; Gagnon J.H. The Social Organization of Sexuality. *Sexual Practices in the United States*, 1994.

[12] Leo Goodman. Snowball Sampling. *Annals of Mathematical Statistics*, 32:148–170, 1961.

[13] Frank Ove. Survey Sampling in Graphs. *Journal of Statistical Planning and Inference*, 1:235–264, 1997a.

[14] Klovdahl A. S. Urban Social Networks: Some Methodological Problems and Possibilities. *The Small World*, 1989.

[15] Marinus Spreen; Moniek Coumans. A Note on Network Sampling in Drug Abuse Research. *Connections,INSNA*, 25, No 1:27–35, 2003.

[16] J.F. French. Pipe Dreams: Crack and the Life in Philadelphia and Newark. *In: Ratner, M.S. (Ed.), Crack Pipe as Pimp. Lexington,New York*, pages 205–232., 1993.

[17] J.A. Inciardi. In Search of the Class Cannon: a Field Study of Professional Pickpockets. *Selected Studies of Crime and Drug Use in Natural*, pages 55–77., 1977.

[18] Walters J. Soloway, I. Workin the Corner: the Ethics and Legality of Ethnographic Fieldwork among Active Heroin Addicts. *Selected Studies of Crime and Drug Use in Natural Settings*, pages 159–178., 1977.

[19] M.H Agar. Ethnography in the Streets and in the Joint. *Street Ethnography: Selected Studies of Crime and Drug Use in Natural Settings. Sage, Beverly Hills,*, pages 143–155., 1977.

[20] Solano Castillo Luis Ramírez Rafael, García Toro Victor. Dando y cogiendo. los puertorriqueños y el deseo homoerótico. *Centro Journal*, XVII, No. I:107–121, 2005.

[21] Biernacki Watters, J.K. Targeted Sampling: Options for the Study of Hidden Populations. *Social Problems*, 36:416–430., 1989.

[22] Sirken Birnbaum, Z.W. Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital Health Stat 2*, No.11, 1965.

[23] Anderson D.W. Kalton, G. Sampling Rare Populations. *Royal Statistic Society*, 149:65–82, 1986.

[24] Ronald Czaja; Cecilia Snowden; Robert Casady. Reporting Bias and Sampling Errors in a Survey of a Rare Population Using Multiplicy Counting Rules. *JASA*, 81:411–419, 1986.

[25] Zacks S. Solomon, H. Optimal Design of Sampling from Finite Populations: a Critical Review and Indication of New Research Areas. *JASA*, 65:653–677., 1970.

[26] Krista Gile Mark S. Handcock. Modeling Social Networks with Sampled or Missing Data. *Center for Statistics and the Social Sciences. University of Washington*, 2002.

[27] Rényi A. Erdös P. "On the Evolution of Random Graphs". *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[28] E.N. Gilbert. "Random Graphs". *Annals of Mathematical Statistics*, 30:1141, 1959.

[29] Woodhouse D.E. Rothenberg R.B. Muth S.Q. Darrow W.W. Muth J.B. Potterat, J.J. and J.U. Reynolds. Aids in Colorado Springs: Is there an Epidemic? *AIDS*, 7:1517–1521, 1993.

APÉNDICES

APÉNDICE A
ALGORITMOS EN R

Algoritmo 1

```
library(sna)
library(network)
gr=rgraph(900,1,0.01,mode="graph")
gr=add.isolates(gr,100)
gr=rmperm(gr)
netw<- network(gr)
gr=as.matrix.network(netw,matrix.type="adjacency")
x=c(1:1000)
u=sample(x,100)
nr=100
propu=length(u)/length(x)
grafo=function(){
s0=sample(x,50)
n1s0=length(intersect(u,s0))
n0s0=length(s0)-n1s0
xc=x[-s0]
smc=gr[s0,xc]
colnames(smc) <-xc
rownames(smc) <-s0
sc=colSums(smc)
y=numeric(length(xc))
for(i in 1:length(xc)){
if(sc[i]!=0) y[i]=c(names(sc[i]))
}
s1=intersect(y,x)
n1s1=length(intersect(s1,u))
n0s1=length(s1)-n1s1
s=c(s0,s1)
s01=intersect(u,s0)
s11=intersect(u,s)
```

```
if(length(s01)==0||length(s11)==0) {s0xs1=integer(0);r22=0;r20=0} else {
s0xs1=gr[s01,s11]
r22=sum(s0xs1)
r20=length(s0xs1)-r22
}
nsc=length(x)-length(s0)-length(s1)
n1s=n1s0+n1s1
n0s=n0s0+n0s1
Res=c(n1s0,n0s0,n1s,n0s,nsc,r22,r20)

a=0.5
b=0.5
g=0.5
h=0.5
AA=n0s+a
BB=n1s+b
GG=r22+g
HH=r20+h
sm1=numeric(nsc+1)
for(i in 1:nsc+1){
sm1[i]=(choose(nsc,i-1)*beta(AA+i-1,nsc+BB-(i-1))*beta(GG,(nsc-i+1)*(n1s0)+HH))
}
SM1=sum(sm1)
if(SM1==0){thetabay=propu;betabay=gden(gr);thebayespr=propu} else {
sm2=numeric(nsc)
for(i in 1:nsc){
sm2[i]=(choose(nsc,i-1)*beta(AA+i,nsc+BB-(i-1))*beta(GG,(nsc-i+1)*(n1s0)+HH))
}
SM2=sum(sm2)
thetabay=SM2/SM1
sm3=numeric(nsc+1)
for(i in 1:nsc+1){
sm3[i]=(choose(nsc,i-1)*beta(AA+i-1,nsc+BB-(i-1))*beta(GG+1,(nsc-i+1)*(n1s0)+HH))
}
SM3=sum(sm3)
betabay=SM3/SM1
sm4=numeric(nsc)
for(i in 1:nsc){
sm4[i]=(nsc)*((choose(nsc-1,i-1)*beta(AA-i+nsc,BB+i))*beta(GG,(i)*(n1s0)+HH))
}
SM4=sum(sm4)
```

```
thebayespr=(n1s+(SM4/SM1))/length(x)
}
thetabay0=1-thetabay
prop=n1s0/length(x)
estim=c(thetabay0,betabay,thebayespr,prop)
estim
}
mm=replicate(nr,grafo())
mediatb=mean(mm[1,])
mediatbpr=mean(mm[3,])
mediab=mean(mm[2,])
medpro=mean(mm[4,])
msmtb=sum((mm[1,]-mediatb)^2)/(nr-1)
msmtbpr=sum((mm[3,]-mediatbpr)^2)/(nr-1)
msmb=sum((mm[2,]-mediab)^2)/(nr-1)
inttb=c(mediatb-(1.96*(msmtb/nr)^(0.5)),mediatb+(1.96*(msmtb/nr)^(0.5)))
inttbpr=c(mediatbpr-(1.96*(msmtbpr/nr)^(0.5)),mediatbpr+(1.96*(msmtbpr/nr)^(0.5)))
intb=c(mediab-(1.96*(msmb/nr)^(0.5)),mediab+(1.96*(msmb/nr)^(0.5)))
estd=rbind(mediatb,mediatbpr,mediab,medpro,msmtb,msmtbpr,msmb)
int=rbind(inttb,inttbpr,intb)
matrix(estd)
int
```

Algoritmo 2

```
library(sna)
library(network)
gr=rgraph(900,1,0.01,mode="graph")
gr=add.isolates(gr,100)
gr=rmperm(gr)
netw<- network(gr)
gr=as.matrix.network(netw,matrix.type="adjacency")
x=c(1:1000)
u=sample(x,100)
#ei:estrato
nr=100
propu=length(u)/length(x)
e=4
e1=sample(x,350)
x1=setdiff(x,e1)
e2=sample(x1,150)
x2=setdiff(x1,e2)
```

```
e3=sample(x2,200)
e4=setdiff(x2,e3)
est=list(e1,e2,e3,e4)
tsampl=c(9,4,5,7)
Us1=list(1,2,3,4)
Us0=list(1,2,3,4)
tn0s0=numeric(4)
tn1s0=numeric(4)

grafo=function(){
for(j in 1:e){
s0=sample(est[[j]],tsampl[j])
n1s0=length(intersect(u,s0))
n0s0=length(s0)-n1s0
xc=x[-s0]
smc=gr[s0,xc]
colnames(smc) <-xc
sc=colSums(smc)
y=numeric(length(xc))
for(i in 1:length(xc)){
if(sc[i]!=0) y[i]=c(names(sc[i]))
}
s1=intersect(y,x)

Us0[[j]]=s0
Us1[[j]]=s1
tn0s0[j]=n0s0
tn1s0[j]=n1s0
}
Tn1s0=sum(tn1s0)
Tn0s0=sum(tn0s0)

U0=c()
U1=c()
for(k in 1:e) {
U0=union(U0,Us0[[k]])
U1=unique(union(U1,Us1[[k]]))
}
s0=U0
Y=U1
S1=setdiff(Y,s0)
```

```
Tn1s1=length(intersect(S1,u))
Tn0s1=length(S1)-Tn1s1
s=c(s0,S1)
s01=intersect(u,s0)
s0s11=intersect(u,s)
if(length(s01)==0||length(s0s11)==0) {s0xs1=integer(0);Tr22=0;Tr20=0} else {
s0xs1=gr[s01,s0s11]
Tr22=sum(s0xs1)
Tr20=length(s0xs1)-Tr22
}
Tnsc=length(x)-length(U0)-length(S1)
Tn1s=Tn1s0+Tn1s1
Tn0s=Tn0s0+Tn0s1
Res=c(Tn1s0,Tn0s0,Tn1s,Tn0s,Tnsc,Tr22,Tr20)
Res

a=0.5
b=0.5
g=0.5
h=0.5

AA=Tn0s+a
BB=Tn1s+b
GG=Tr22+g
HH=Tr20+h

sm1=numeric(Tnsc+1)
for(i in 1:Tnsc+1){
sm1[i]=(choose(Tnsc,i-1)*beta(AA+i-1,Tnsc+BB-i+1)*beta(GG,(Tnsc-i+1)*Tn1s0+HH))
}
SM1=sum(sm1)
if(SM1==0){thetabay=propu;betabay=gden(gr);thebayespr=propu} else {
sm2=numeric(Tnsc+1)
for(i in 1:Tnsc+1){
sm2[i]=(choose(Tnsc,i-1)*beta(AA+i,Tnsc+BB-i+1)*beta(GG,(Tnsc-i+1)*Tn1s0+HH))
}
SM2=sum(sm2)
thetabay=SM2/SM1
sm3=numeric(Tnsc+1)
for(i in 1:Tnsc+1){
sm3[i]=(choose(Tnsc,i-1)*beta(AA+i-1,Tnsc+BB-i+1)*beta(GG+1,(Tnsc-i+1)*Tn1s0+HH))
```

```
}
SM3=sum(sm3)
betabay=SM3/SM1
sm4=numeric(Tnsc)
for(i in 1:Tnsc){
sm4[i]=(Tnsc)*((choose(Tnsc-1,i-1)*beta(AA-i+Tnsc,BB+i))*beta(GG,(i)*Tn1s0+HH))
}
SM4=sum(sm4)
thebayespr=(Tn1s+(SM4/SM1))/length(x)
}
thetabay0=1-thetabay
betabay
prop=Tn1s0/length(x)
estim=c(thetabay0,betabay,thebayespr,prop)
estim
}
mm=replicate(nr,grafo())
mediatb=mean(mm[1,])
mediatbpr=mean(mm[3,])
mediab=mean(mm[2,])
medpro=mean(mm[4,])
msmtb=sum((mm[1,]-mediatb)^2)/(nr-1)
msmtbpr=sum((mm[3,]-mediatbpr)^2)/(nr-1)
msmb=sum((mm[2,]-mediab)^2)/(nr-1)
inttb=c(mediatb-(1.96*(msmtb/nr)^(0.5)),mediatb+(1.96*(msmtb/nr)^(0.5)))
inttbpr=c(mediatbpr-(1.96*(msmtbpr/nr)^(0.5)),mediatbpr+(1.96*(msmtbpr/nr)^(0.5)))
intb=c(mediab-(1.96*(msmb/nr)^(0.5)),mediab+(1.96*(msmb/nr)^(0.5)))
estd=rbind(mediatb,mediatbpr,mediab,medpro,msmtb,msmtbpr,msmb)
int=rbind(inttb,inttbpr,intb)
matrix(estd)
int
nr
```

Algoritmo 3

```
gr=rgraph(900,1,0.01,mode="graph")
gr=add.isolates(gr,100)
gr=rmperm(gr)
netw<- network(gr)
gr=as.matrix.network(netw,matrix.type="adjacency")
x=c(1:1000)
u=sample(x,100)
#ei:estrato
```

```
nr=100
e=4
e1=sample(x,350)
x1=setdiff(x,e1)
e2=sample(x1,150)
x2=setdiff(x1,e2)
e3=sample(x2,200)
e4=setdiff(x2,e3)
propu=length(u)/length(x)
est=list(e1,e2,e3,e4)
tsampl=c(9,4,5,7)
tn1s=numeric(4)
tn0s=numeric(4)
tn0s0=numeric(4)
tn1s0=numeric(4)
tr20=numeric(4)
tr22=numeric(4)
Us1=list(1,2,3,4)

grafo=function(){
for(j in 1:e){
s0=sample(est[[j]],tsampl[j])
n1s0=length(intersect(u,s0))
n0s0=length(s0)-n1s0
xc=setdiff(est[[j]],s0)
smc=gr[s0,xc]
colnames(smc) <-xc
sc=colSums(smc)
y=numeric(length(xc))
for(i in 1:length(xc)){
if(sc[i]!=0) y[i]=c(names(sc[i]))
}
s1=intersect(y,x)
n1s1=length(intersect(s1,u))
n0s1=length(s1)-n1s1
s=c(s0,s1)
s01=intersect(u,s0)
s11=intersect(u,s)
if(length(s01)==0||length(s11)==0) {s0xs1=integer(0);r22=0;r20=0} else {
s0xs1=gr[s01,s11]
r22=sum(s0xs1)
```

```
r20=length(s0xs1)-r22
}
n1s=n1s0+n1s1
n0s=n0s0+n0s1

Us1[[j]]=s1
tn0s0[j]=n0s0
tn1s[j]=n1s
tn0s[j]=n0s
tn1s0[j]=n1s0
tr22[j]=r22
tr20[j]=r20

}

U1=c()
for(k in 1:e) {
U1=union(U1,Us1[[k]])
}

tnsc=length(x)-sum(tsampl)-length(U1)
Res=c(sum(tn1s0),sum(tn0s0),sum(tn1s),sum(tn0s),tnsc,sum(tr22),sum(tr20))



stn1s0=sum(tn1s0)
stn0s0=sum(tn0s0)
stn1s=sum(tn1s)
stn0s=sum(tn0s)
str22=sum(tr22)
str20=sum(tr20)

a=0.5
b=0.5
g=0.5
h=0.5

AA=stn0s+a
BB=stn1s+b
GG=str22+g
HH=str20+h
sm1=numeric(tnsc+1)
```

```
for(i in 1:tnsc+1){
sm1[i]=(choose(tnsc,i-1)*beta(AA+i-1,tnsc+BB-i+1)*beta(GG,(tnsc-i+1)*stn1s0+HH))
}
SM1=sum(sm1)
if(SM1==0){thetabay=propu;betabay=gden(gr);thebayespr=propu} else {
sm2=numeric(tnsc+1)
for(i in 1:tnsc+1){
sm2[i]=(choose(tnsc,i-1)*beta(AA+i,tnsc+BB-i+1)*beta(GG,(tnsc-i+1)*stn1s0+HH))
}
SM2=sum(sm2)
thetabay=SM2/SM1
sm3=numeric(tnsc+1)
for(i in 1:tnsc+1){
sm3[i]=(choose(tnsc,i-1)*beta(AA+i-1,tnsc+BB-i+1)*beta(GG+1,(tnsc-i+1)*stn1s0+HH))
}
SM3=sum(sm3)
betabay=SM3/SM1
sm4=numeric(tnsc+1)
for(i in 1:tnsc+1){
sm4[i]=(tnsc)*(choose(tnsc-1,i-1)*beta(AA-i+tnsc,BB+i)*beta(GG,(i)*stn1s0+HH))
}
SM4=sum(sm4)
prop=stn1s0/length(x)
thebayespr=(stn1s+(SM4/SM1))/length(x)
}
thetabay0=1-thetabay
estim=c(thetabay0,betabay,thebayespr,prop)
estim
}
mm=replicate(nr,grafo())
mediatb=mean(mm[1,])
mediab=mean(mm[2,])
mediatbpr=mean(mm[3,])
medpro=mean(mm[4,])
msmtb=sum((mm[1,]-mediatb)^2)/(nr-1)
msmtbpr=sum((mm[3,]-mediatbpr)^2)/(nr-1)
msmb=sum((mm[2,]-mediab)^2)/(nr-1)
inttb=c(mediatb-(1.96*(msmtb/nr)^(0.5)),mediatb+(1.96*(msmtb/nr)^(0.5)))
inttbpr=c(mediatbpr-(1.96*(msmtbpr/nr)^(0.5)),mediatbpr+(1.96*(msmtbpr/nr)^(0.5)))
intb=c(mediab-(1.96*(msmb/nr)^(0.5)),mediab+(1.96*(msmb/nr)^(0.5)))
```

```
estd=rbind(mediatb,mediatbpr,mediab,medpro,msmtb,msmtbpr,msmb)
int=rbind(inttb,inttbpr,intb)
matrix(estd)
int
nr
```

Algoritmo 4

```
gr=rgraph(900,1,0.01,mode="graph")
gr=add.isolates(gr,100)
gr=rmperm(gr)
netw<- network(gr)
gr=as.matrix.network(netw,matrix.type="adjacency")
x=c(1:1000)
u=sample(x,100)
#ei:estrato
nr=100
e=4

e1=sample(x,350)
x1=setdiff(x,e1)
e2=sample(x1,150)
x2=setdiff(x1,e2)
e3=sample(x2,200)
e4=setdiff(x2,e3)
propu=length(u)/length(x)
te=c(350,150,200,300)
Us0=list(1,2,3,4)
est=list(e1,e2,e3,e4)
tsampl=c(9,4,5,7)
m=matrix(nrow = 4, ncol=3)
grafo=function(){
for(j in 1:e){
s0=sample(est[[j]],tsampl[j])
n1s0=length(intersect(u,s0))
n0s0=length(s0)-n1s0
xc=x[-s0]
smc=gr[s0,xc]
colnames(smc) <-xc
rownames(smc) <-s0
sc=colSums(smc)
y=numeric(length(xc))
for(i in 1:length(xc)){
```

```
if(sc[i]!=0) y[i]=c(names(sc[i]))
}
s1=intersect(y,x)
Us0[[j]]=s0
n1s1=length(intersect(s1,u))
n0s1=length(s1)-n1s1
s=c(s0,s1)
s01=intersect(u,s0)
s11=intersect(u,s)
if(length(s01)==0||length(s11)==0) {s0xs1=integer(0);r22=0;r20=0} else {
s0xs1=gr[s01,s11]
r22=sum(s0xs1)
r20=length(s0xs1)-r22
}
nsc=length(x)-length(s0)-length(s1)
n1s=n1s0+n1s1
n0s=n0s0+n0s1
Res=c(n1s0,n0s0,n1s,n0s,nsc,r22,r20)

a=0.5
b=0.5
g=0.5
h=0.5
AA=n0s+a
BB=n1s+b
GG=r22+g
HH=r20+h
sm1=numeric(nsc+1)
for(i in 1:nsc+1){
sm1[i]=(choose(nsc,i-1)*beta(AA+i-1,nsc+BB-(i-1))*beta(GG,(nsc-i+1)*n1s0+HH))
}
SM1=sum(sm1)
if(SM1==0){thetabay=propu;betabay=gden(gr);thebayespr=propu} else {
sm2=numeric(nsc+1)
for(i in 1:nsc+1){
sm2[i]=(choose(nsc,i-1)*beta(AA+i,nsc+BB-(i-1))*beta(GG,(nsc-i+1)*n1s0+HH))
}
SM2=sum(sm2)
thetabay=SM2/SM1
sm3=numeric(nsc+1)
for(i in 1:nsc+1){
```

```
sm3[i]=(choose(nsc,i-1)*beta(AA+i-1,nsc+BB-(i-1))*beta(GG+1,(nsc-i+1)*n1s0+HH))
}
SM3=sum(sm3)
betabay=SM3/SM1
sm4=numeric(nsc)
for(i in 1:nsc){
sm4[i]=(nsc)*((choose(nsc-1,i-1)*beta(AA-i+nsc,BB+i))*beta(GG,(i)*n1s0+HH))
}
SM4=sum(sm4)
thebayespr=(n1s+(SM4/SM1))/length(x)
}
thetabay0=1-thetabay
prop=n1s0/length(x)
m[j,1]=thetabay0
m[j,2]=thebayespr
m[j,3]=betabay
final=c(sum(m[,1]*te)/length(x),sum(m[,2]*te)/length(x),sum(m[,3]*te)/length(x))
}
final
}
mm=replicate(nr,grafo())
mediatb=mean(mm[1,])
mediatbpr=mean(mm[2,])
mediab=mean(mm[3,])
msmtpr=(sum((mm[2,]-mediatbpr)^2))/(nr-1)
msmt=(sum((mm[1,]-mediatb)^2))/(nr-1)
msmb=(sum((mm[3,]-mediab)^2))/(nr-1)
inttb=c(mediatb-(1.96*(msmtb/nr)^(0.5)),mediatb+(1.96*(msmtb/nr)^(0.5)))
inttbpr=c(mediatbpr-(1.96*(msmtbpr/nr)^(0.5)),mediatbpr+(1.96*(msmtbpr/nr)^(0.5)))
intb=c(mediab-(1.96*(msmb/nr)^(0.5)),mediab+(1.96*(msmb/nr)^(0.5)))

estd=rbind(mediatb,mediatbpr,mediab,medpro,msmtb,msmtbpr,msmb)
int=rbind(inttb,inttbpr,intb)
matrix(estd)
int
nr
```