

Universidad Mayor de San Andrés
Facultad de Ciencias Puras y Naturales
Carrera de Estadística



Trabajo Dirigido

Sistema De Muestreo RDS

Aplicación en un Estudio de Conocimientos, Actitudes y Comportamientos asociados al VIH/SIDA en población de Hombres que tienen Sexo con Hombres (HSH)

Postulante: Wilson René Alarcon Flores

Docente Tutor: Lic. Rubén Belmonte Coloma

La Paz, Bolivia

2012

Contenido

Índice de figuras	iv
Índice de tablas	v
Glosario de términos y abreviaturas	vi
Notación usada en este documento	vii
Resumen.....	viii
1.1. Introducción	1
1.2. Planteamiento del problema	2
1.3. Objetivos	3
1.3.1. Objetivo General	3
1.3.2. Objetivos Específicos	3
1.4. Alcances y limitaciones	3
1.4.1. Alcances.....	3
1.4.2. Limitaciones	3
2.1. Fundamentos de la investigación.....	4
2.2. Sistema de Muestreo RDS.....	4
2.2.1. Descripción	4
2.2.2. Poblaciones ocultas.....	6
2.2.3. Redes sociales	7
2.2.3.1. Redes estructuradas de tipo centro periferia	9
2.2.3.2. Redes estructuradas en cohorte	9
2.2.3.3. Redes estructuradas bipartitas	9
2.2.4. Tipos de muestreo aplicables a estas poblaciones	9
2.2.4.1. Muestreo Bola de Nieve (snow-ball).....	9
2.2.4.2. Muestreo basado en establecimientos (Facility-Based)	10
2.2.4.3. Muestreo Dirigido	11
2.2.4.4. Muestreo por lugar y tiempo (Time-Location Sampling - TLS).....	11
2.2.4.5. Muestreo convencional por conglomerados	12
2.3. El trabajo con RDS	13
2.3.1. Muestreo RDS como un proceso de Markov	16
2.3.1.1. RDS, un proceso de cadenas de Markov mediante métodos de Monte Carlo (MCMC)	22

2.3.1.2.	Aplicación a RDS	24
2.3.2.	Modelo de homofilia, evaluación del sesgo	26
2.4.	Estimaciones poblacionales	33
2.4.1.	Uso de las redes sociales para realizar estimaciones hacia la población.....	33
2.4.2.	Uso de la muestra para realizar estimaciones acerca de las redes sociales	40
2.4.2.1.	Supuestos	40
2.4.2.2.	Consecuencias de estos supuestos	42
2.4.3.	Construyendo los estimadores.....	43
2.4.3.1.	Estimación de las probabilidades de selección recíprocas (S_{AB} y S_{BA}).....	43
2.4.3.2.	Estimación del grado medio del grupo (D_A y D_B).....	44
2.4.3.3.	Estimación de la proporción poblacional (P).....	47
2.4.4.	Estimación de la varianza	48
2.4.5.	Medición de la precisión de la estimación	50
2.4.5.1.	Intervalos de confianza	50
2.4.5.2.	Procedimiento de autoreposición.....	53
2.4.5.3.	Cálculo del tamaño muestral	55
2.4.5.4.	Efecto de diseño	56
3.1.	Introducción	57
3.2.	Características de la muestra y selección de los entrevistados	58
3.2.1.	Cálculo del tamaño de la muestra.....	58
3.2.2.	Descripción de las semillas y el proceso de reclutamiento.....	60
3.2.3.	Característica de la muestra.....	63
3.2.4.	Caracterización de las redes sociales	64
3.3.	Estimación RDS.....	70
3.3.1.	Detalles de la estimación de RDS	70
3.3.1.1.	Matriz de reclutamiento	74
3.3.1.2.	Proporción Muestral	74
3.3.1.3.	Proporción poblacional	75
3.3.1.4.	Equilibrio de la muestra	79
3.3.1.5.	Reclutamiento diferencial	80
3.3.1.6.	Suavización de los datos.....	81
3.3.1.7.	Homofilia	81
3.3.1.8.	Estimación de la varianza	84

3.3.1.9.	Intervalos de confianza	86
3.3.1.10.	Sesgo asociado a las estimaciones de proporción de la muestra (S)	86
3.3.1.11.	Explorando las diferencias entre la muestra y las estimaciones de población	87
4.1.	Conclusiones.....	90
4.2.	Recomendaciones	91
	Carta de conformidad de la institución.....	93
	Bibliografía	94
4.3.	Boleta de encuesta	95
4.4.	Software RDSAT y manual (versión digital).....	104

Índice de figuras

Figura 1. Redes Sociales	7
Figura 2. Proceso de estimación RDS	33
Figura 3 - Proceso de inferencia estadística.....	54
Figura 4 - Casos de VIH/SIDA 1984 - 2011.....	57
Figura 5- Distribución del grado, orientación sexual	62
Figura 6 - Distribución del grado, trabajo sexual	62
Figura 7 - Patrón de reclutamiento del estudio	65
Figura 8- Estructura de las redes sociales, orientación sexual.....	66
Figura 9 - Estructura de las redes sociales según trabajo sexual	66
Figura 10 –Número de entrevistados en cada una de las olas	67
Figura 11 - Evolución del % de entrevistados según orientación sexual	68
Figura 12 - Evolución del % de entrevistados según comportamiento de trabajo sexual (n=373, incluye semillas)	69
Figura 13 - Intervalo de confianza (95%) para el grado no ajustado, orientación sexual.....	72
Figura 14 - Intervalo de confianza (95%) para el grado no ajustado, trabajo sexual.....	73
Figura 15 - Probabilidades de transición entre grupos e intragrupos, orientación sexual	74
Figura 16 - Probabilidades de transición entre grupos e intragrupos, trabajo sexual	75
Figura 17 – Proporción poblacional, orientación sexual.....	76
Figura 18 – Proporción poblacional, trabajo sexual.....	76
Figura 19 – Proporción poblacional, edad	77
Figura 20- Desplazamiento de olas de reclutados	78
Figura 21 - Homofilia, orientación sexual	83
Figura 22 - Homofilia, trabajo sexual	83
Figura 23 - Homofilia, edad	84

Índice de tablas

Tabla 1 - Características de las semillas participantes en el estudio (n=17)	61
Tabla 2 – Características socio demográficas de los entrevistados (n=356 no incluye semillas)	64
Tabla 3 - Distribución de los entrevistados, orientación sexual (n=373, incluye semillas).....	68
Tabla 4 - Distribución de los entrevistados según trabajo sexual (incluye semillas) (n=373, incluye semillas).....	69
Tabla 5 -Estimadores RDS, variable en estudio orientación sexual	71
Tabla 6 - Intervalos de confianza para las proporciones poblacionales P_a , P_b , P_c (alfa=0.05) – Orientación sexual.....	71
Tabla 7 - Estimadores RDS, variable en estudio trabajo sexual	72
Tabla 8 - Intervalos de confianza para las proporciones poblacionales P_a y P_b (alfa=0.05) – trabajo sexual	73
Tabla 9 - Matriz de reclutamiento diferencial.....	80
Tabla 10 - Comparación de datos, orientación sexual	81
Tabla 11 - Comparación de datos, trabajo sexual.....	81
Tabla 12 - Remuestreos de los intervalos de confianza para P, orientación sexual	86
Tabla 13 - Remuestreos de los intervalos de confianza para P, trabajo sexual	86
Tabla 14 - Diferencias entre muestra y población, orientación sexual.....	88
Tabla 15 - Diferencias entre muestra y población, trabajo sexual.....	88

Glosario de términos y abreviaturas

Homofilia o endogamia	Mide las asociaciones de las redes sociales
HSH	Hombre que tiene sexo con hombres
GBT	Gay, bisexual, trans
heterofilia	cuando todos los lazos se forman fuera del grupo
Profundidad sociométrica	Cantidad de relaciones sociales que tienen las personas pertenecientes a un grupo que permite que el muestreo por referencia en cadena avance
TS	Trabajadora Sexual
Grado	Número de personas que pueden ser reclutadas

Notación usada en este documento

U	es el conjunto de todos los individuos en la población
L	es el conjunto de todos los individuos en la muestra
E	Distribución de equilibrio
P	Proporción muestral
H	Homofilia
S	Probabilidad de selección
S_{AB}	Es la probabilidad estimada de que alguien de la serie A seleccione a alguien de la serie B
R_A	Suma del grado de todas las personas en el grupo A
N_A	Número de personas en el grupo A
$p_A(D)$	Distribución del grado de la población
p_i	Probabilidad que una persona sea seleccionada en una selección específica
A, B	son conjuntos disjuntos de individuos
N_x	es el número de unidades de muestra de X Conjunto
P_A, P_B	son las proporciones de la población de cada tipo, A, B, etc
T_{AB}	es el número de selecciones del Grupo A al Grupo B
T_A	es el número total de veces las personas de tipo A son reclutados
R_A	número de amistades que irradian del grupo A
d_i	es el grado del individuo i
d_x	es el grado medio de individuos del conjunto X

Resumen

El objetivo de este trabajo es mostrar la aplicación del sistema de muestreo Respondent-Driven Samplig, o Muestreo Dirigido por el Encuestado, método de muestreo orientado a investigar en poblaciones ocultas.

Poblaciones ocultas son aquellas que no tienen un marco muestral y donde el nivel de estigma hace que las personas pertenecientes a esta población no se identifiquen y/o no quieran participar en estudios convencionales.

Este sistema de muestreo corrige las deficiencias de muestreos no probabilísticos, comúnmente usados en este tipo de poblaciones, estos sistemas de muestreo no probabilísticos generalmente eran usados de manera meramente referencial al no permitir la construcción de estimadores y por lo tanto no poder realizar la inferencia estadística.

El método de muestreo RDS se aleja un tanto de la estadística tradicional, abordando poblaciones que tienen diferentes características a las regulares, a través de condiciones innovadoras que permitan generar un espacio de análisis estadístico formal, aproximándose a los estimadores tradicionales para enmarcarse dentro de los parámetros de análisis regular.

La aplicación del RDS se realizó en un estudio desarrollado en la ciudad de Santa Cruz, a hombres que tienen sexo con hombres que fueron identificados por sus pares como parte del grupo y cumplieron ciertas características que les permitió ser parte del estudio. Se eligió esta ciudad para la investigación por el alto índice de prevalencia que presenta en casos de VIH, en el grupo de investigación determinado (HSH).

La metodología se basó en la selección de un grupo inicial denominado *semillas* que fueron los responsables de generar el resto de la muestra, la misma que se estructuró en forma de árbol a partir de las semillas y creció de manera geométrica bajo el método de referencia en cadena, a través de 10 olas continuas de reclutamiento que plasmó una muestra de 356 personas encuestadas, tamaño muestral que incluye a las semillas.

El proceso demostró que se alcanzó el equilibrio muestral, es decir, la estructura de la muestra es independiente de las semillas, que es una de las principales preocupaciones de este sistema de muestreo, la estructura muestral posterior al equilibrio se mantuvo durante toda la encuesta.

El proceso de análisis se concentró en dos variables consideradas como representativas de toda la estructura, la primera, orientación sexual, por ser una variable de composición poblacional que evidencia cómo está conformada la *estructura* poblacional bajo estudio. La segunda variable es el trabajo sexual, elegida por ser representativa del *comportamiento* del grupo en relación con el VIH.

Se encontró además, que las estructuras de las redes son muy diferentes al interior de los subgrupos presentes por orientación sexual, con niveles de homofilia que demuestran un ligero nivel de endogamia para los sub grupos de gays y travestis y una heterofilia en el subgrupo de bisexuales, estos procesos demuestran la forma en que se estructuran las redes sociales.

Se encontró que los procesos de reclutamiento son altamente dependientes del conocimiento del contexto en el que se desarrolló la encuesta, proceso que permitió predecir *a priori* el comportamiento y la profundidad de las redes sociales.

Estas redes sociales han mostrado profundidades sociométricas adecuadas, que proporcionaron un contexto apropiado para el desarrollo de la metodología, además de permitir un equilibrio muestral bastante rápido, permitieron que los estimadores muestrales sean útiles para la estimación de la población en estudio. Este método de muestreo realiza el proceso de estimación, primero a la red social y posteriormente a la población que es una diferencia esencial con otros sistemas de muestreo.

El desarrollo de todo el procedimiento demostró que el método de investigación es bastante adecuado para este tipo de poblaciones y permitió encontrar los estimadores para la población, que se considera son insesgados o al menos tienen un sesgo mínimo, dado que son independientes de la forma en que se seleccionó la muestra.

Capítulo 1: Introducción

1.1. Introducción

La existencia de poblaciones, tales como, usuarios de drogas, trabajadoras(es) sexuales, personas viviendo con VIH, presentan varios problemas para el desarrollo de estudios utilizando sistemas de muestreo tradicionales (aleatorio, sistemático, conglomerados, etc.), estos problemas están asociados a la falta de marcos muestrales existentes, dado que los límites de estas poblaciones son desconocidos, se debe considerar también que estas poblaciones tienen características de estigmatización y quieren mantener la privacidad, por lo tanto las repuestas a estudios tradicionales carecerán de sinceridad para proteger su privacidad.

Todas las características descritas hacen que estas poblaciones se consideren como “**poblaciones ocultas**”, por lo tanto, se hace necesaria la aplicación de un método de muestreo probabilístico no tradicional, que se presenta en este trabajo, que dé, como resultado estimaciones asintóticamente insesgadas, este sistema de muestreo es Respondent-Driven Samplig, **RDS**, o Muestreo Dirigido por el Encuestado (traducción aproximada, no existe una traducción correcta del término por lo que en este documento se usará el acrónimo en inglés).

Este método de muestreo, corrige los problemas derivados de la falta de representatividad que limitan la validez de los muestreos no probabilísticos, tales como los métodos de bola de nieve (snowball sampling)¹, el muestreo a través de informantes clave (key informant sampling)² o a partir del uso de muestreos dirigidos (targeted sampling)³.

A diferencia de otros métodos de referencia en cadena, el RDS permite la evaluación de las probabilidades de inclusión de los miembros de la población bajo estudio, basado en un modelo matemático del proceso de reclutamiento. Este modelo se deriva de una síntesis y una extensión de la teoría de las cadenas de Markov y provee la base para el cálculo de los estimadores, los errores estándar y los intervalos de confianza.

Las estimaciones hacia la población tendrán una diferencia marcada con los métodos tradicionales de muestreo, esta diferencia radica en que, con este método no se puede realizar directamente la inferencia de la muestra a la población, sino más bien se realizará una estimación a la red social y de esta a la población que cumple con la característica de ser HSH, que estadísticamente podría ser denominada pseudo población, sin embargo socialmente no es aceptable el término, por lo que se la seguirá denominando población en este estudio.

La aplicación práctica desarrolla una investigación entre hombres que tienen sexo con hombres, en la ciudad de Santa Cruz, indagando acerca de sus conocimientos, actitudes y comportamientos relacionados al VIH/SIDA.

¹ Goodman (1961)

² Deux y Callaghan, (1985)

³ Watters y Biernacki, (1989)

La importancia de este tipo de estudios en esta población específica, radica en la necesidad de tener información clara en cuanto a los componentes de la epidemia del VIH en el país, sabiendo que la epidemia en Bolivia está catalogada como una epidemia concentrada:

“En las epidemias concentradas, en los casos en que el 5% o más de algún grupo de población de más alto riesgo se encuentra infectado por el VIH (por ejemplo, usuarios de drogas inyectables, trabajadores sexuales u hombres que tienen sexo con hombres), la vigilancia se lleva a cabo entre esos grupos. Se presta atención especial a los comportamientos que sirven de conexión entre esos grupos con comportamientos de riesgo y la población general.”⁴

Y la mayor prevalencia se encuentra en la ciudad de Santa Cruz.

1.2. Planteamiento del problema

La identificación de conocimientos y comportamientos asociados al VIH/SIDA en redes sociales de población de Hombres que tienen Sexo con Hombres (HSH), a través del sistema de muestreo dirigido por el entrevistado RDS.

Esta investigación centrará su atención en la aplicación del sistema de Respondent-Driven Sampling, RDS, que es un muestreo de tipo probabilístico que permite realizar estimaciones insesgadas en poblaciones ocultas y no está limitado a los miembros del grupo encuestado.

Este método de muestreo corrige los problemas derivados por la falta de representatividad que limitan la validez de los muestreos no probabilísticos, usados comúnmente en las poblaciones de difícil alcance, ocultas o semi ocultas, que son aquellas que no poseen un marco muestral, este está incompleto o no es representativo de la población real. Los estimadores poblacionales generados por el RDS son asintóticamente insesgados⁵ cuando ciertos supuestos son satisfechos, sin importar cómo se seleccionan las semillas iniciales.

La aplicación se hace válida dado que actualmente el VIH/SIDA se ha convertido en un problema de salud pública, por lo tanto es muy importante tener un sistema de vigilancia para el análisis de la epidemia, este estudio está destinado a conocer tanto los conocimientos como los comportamientos relacionados a la transmisión del VIH/SIDA en la población en hombres que tienen sexo con hombres (HSH).

El ámbito geográfico determinado para la investigación es la ciudad de Santa Cruz, por ser la ciudad en Bolivia con más alta prevalencia en VIH.

Para esta investigación, la población de HSH es considerada como población oculta dado que no existe un marco muestral de referencia, el tamaño poblacional y los márgenes reales de esta población son desconocidos, por lo que la aplicación de sistemas de muestreo tradicional no son útiles, adicionalmente a esto se debe agregar que la población bajo análisis es estigmatizada y no puede ser reconocida por personas ajenas a los grupos.

⁴ ONUSIDA (2000). Programas Nacionales de SIDA, Guía para el monitoreo y la evaluación. *ONUSIDA/00.17E (Original: inglés, junio de 2000).*

⁵ Estimador Asintóticamente Insesgado, significa que al aumentar el tamaño de la muestra, su media tiende a coincidir con el parámetro θ , y por lo tanto, su sesgo tiende a cero.

$$\text{Esto es } \lim_{n \rightarrow \infty} E[\hat{\theta}] = \theta$$

En el presente documento la insesgadez de los estimadores será mostrada para RDS como una extensión de las cadenas de Markov.

Por lo tanto el método de muestreo propuesto es aplicable en este contexto y permite un análisis de las redes sociales en esta población.

1.3. Objetivos

1.3.1. Objetivo General

Desarrollo de la metodología de RDS como método de muestreo aplicable a poblaciones ocultas, que sea probabilístico que genere estimadores insesgados de la población objetivo, estudiando los comportamientos y prácticas relacionadas al VIH de hombres que tienen sexo con hombres

1.3.2. Objetivos Específicos

- Identificar el universo de inferencia, en este caso la identificación de redes sociales de gays, bisexuales y trans;
- Extender el universo a conjuntos mayores, en el caso: alcanzar a personas no vinculadas a las redes tradicionales ya conocidas
- Mostrar que las estimaciones son asintóticamente insesgadas independiente de la selección de semillas
- Medir el índice de homofilia para medir la extensión de los vínculos existentes dentro del grupo.

1.4. Alcances y limitaciones

1.4.1. Alcances

El alcance del presente trabajo es la investigación de comportamientos, con un planteamiento muestral que permita la comparación posterior a las intervenciones que se realizan con el grupo poblacional bajo estudio que generen estimaciones asintóticamente insesgadas.

1.4.2. Limitaciones

Al no conocer los parámetros reales de la población, el insesgamiento que se prueba será por aproximación de cadenas de markov y de hansen-horwitz, y se asumirá que los valores conseguidos en el estudio cumplen con ser estimadores insesgados.

Para este documento se asume un reclutamiento uniforme, este supuesto de reclutamiento uniforme es casi seguro que no se cumplen en la mayoría de los estudios a causa de sesgo de selección en las decisiones de los reclutadores a quién recurrir y de los reclutas sobre la posibilidad de participar.

Capítulo 2: Marco Teórico

2.1. Fundamentos de la investigación

Este método de muestreo prueba que las posibilidades ofrecidas por los métodos basados en análisis de redes se pueden combinar con la validez estadística de los métodos estándar de muestreo probabilístico. En este sentido, RDS se presenta como una sustentación matemática válida para la investigación en poblaciones ocultas. Se trata de mostrar su validez para ser utilizado con aquellas poblaciones que no tienen registros en marcos muestrales y también puede ser utilizado en la investigación de subpoblaciones de personas con comportamientos particulares y para las cuales el costo de construir un marco muestral sería muy elevado y por lo tanto la información no puede ser recopilada por instrumentos tradicionales.

Un primer reto, relativo a este caso en particular, para la vigilancia de poblaciones de alto riesgo de contraer el VIH, es obtener muestras representativas que permitan medir la vigilancia, por lo tanto es necesario aplicar un método factible y capaz de producir estimaciones que se espera tengan un sesgo mínimo y pueda ser aplicado en subgrupos poblacionales donde los modelos tradicionales de muestreo no son eficientes en una estrategia de vigilancia. Por lo tanto se considera el método de muestreo RDS, como una aproximación a métodos de segunda generación para la vigilancia epidemiológica.

2.2. Sistema de Muestreo RDS

2.2.1. Descripción

Respondent-Driven Sampling (RDS) es un riguroso sistema de muestreo que utiliza la referencia en cadena que permite la inferencia estadística de la población objetivo mediante el control de las fuentes de sesgo generalmente asociadas con este tipo de muestreo.

RDS se aplica actualmente en los EE.UU. y en todo el mundo para estudiar poblaciones difíciles de alcanzar u ocultas.

Los Centros para el Control y Prevención para Enfermedades utilizan la metodología RDS para supervisar la conducta de riesgo de VIH entre consumidores de drogas inyectables en 25 ciudades en los EE.UU. y Family Health International, la mayor organización sin fines de lucro en materia de salud pública mundial, lo está utilizando en más de una docena de países⁶.

La principal ventaja del sistema RDS es que no requiere un marco de muestreo. Por lo tanto, es eficaz para poblaciones estigmatizadas, ocultas o de difícil acceso, donde las organizaciones u instituciones no aglutinan a todas las personas de la población y sus afiliados no son necesariamente representativos de toda la población.

⁶ Lang 2004; Heckathorn (2002)

El sistema de referencia en cadena se diferencia de los muestreos ordinarios porque los encuestados están unidos entre sí por un vínculo de reclutamientos.

El RDS proporciona probabilidades de muestreo válidas, a través de la construcción de la matriz de reclutamiento, que asigna las probabilidades de selección a los grupos de la población incluidos en el grupo.

Como ya se mencionó, una de las ventajas del sistema RDS es que no requiere de un exhaustivo proceso previo para la construcción de marcos de muestreo. Con RDS, el marco muestral para otras investigaciones subsecuentes es construido durante el proceso de toma de muestra, al reclutar a sus pares y documentar sistemáticamente los métodos de reclutamiento. Otra ventaja del sistema RDS es que se basa en un doble sistema de incentivos, la recompensa o premio en combinación con la presión del grupo, porque los que no participarían por la recompensa lo pueden hacer como favor a un amigo o compañero y con esto se puede esperar se reduzca el sesgo por no respuesta.

El proceso se inicia con la selección de un grupo de personas denominadas *semillas*, que representan la ola 0 del estudio, estas semillas serán las encargadas de reclutar a personas que cumplan ciertas características para ser entrevistadas que representan la ola 1, quienes a su vez proveerán referencias de otra personas que puedan ser entrevistadas representando las nuevas olas, este flujo se realizará hasta alcanzar el tamaño de muestra requerido.

Este método tiene validez externa ya que no se limita a los miembros accesibles del subgrupo, sino que extiende la posibilidad de muestrear a todos los integrantes objeto de la vigilancia epidemiológica, motivo principal para la realización de este tipo de estudio de comportamiento.

Con el método RDS las *semillas* son enlistadas como reclutadores temporales, estas semillas reciben una explicación del estudio y un número limitado de cupones que pueden ser utilizados para reclutar a un compañero que sea elegible para el estudio. La semilla refiere a sus pares para el estudio, proporcionándoles un cupón que tiene un número de serie único. Si su par cumple con las características y se inscribe en el estudio, la semilla, es elegible para un *premio*⁷ por sus esfuerzos de reclutamiento.

Cada entrevistado referido recibe un número similar de cupones (tres o cuatro), hasta que se cumpla el tamaño de la muestra. Debido a que los entrevistados son referidos, deben presentarse en el sitio de estudio, el reclutamiento es totalmente voluntario, por lo que los entrevistadores no necesitan los nombres o datos de contacto de los posibles participantes.

Entre las características principales que distinguen al sistema de muestreo RDS, es que las *semillas* son limitadas en el número de entrevistados que pueden reclutar por el número limitado de cupones que reciben, reduciendo así al mínimo la influencia de las semillas iniciales en la composición de la muestra final. Limitar el número de reclutas por persona alienta largas cadenas de reclutamiento, lo que aumenta el "alcance" de la muestra en más "sitios" ocultos de la población.

⁷ Para este estudio se previó como método de premiación la consulta médica en centros especializados además de condones masculinos

Otra característica es que la relación entre los reclutadores y reclutas es documentada, de manera que los sesgos de selección pueden ser evaluados y ajustados en la etapa de análisis, la información sobre el tamaño de la red personal de cada encuestado se recoge para permitir el análisis ponderado a través de post-estratificación para compensar el sobre muestreo de los encuestados con grandes redes sociales. Por ejemplo, un encuestado RDS típico, **A**, refiere a **B**, sin saber el nombre de forma individual, preguntamos a **B** por su relación con la persona que le dio el cupón (en este caso **A**), **B** responde - *pareja sexual ocasional*. A través del número de serie de los cupones se obtiene el vínculo con **A** con la información de relación obtenida en la encuesta. Cuando **A** vuelve para recoger el premio, se pide su relación con la persona que él refirió (en este caso **B**) y estos datos se convierten en información vinculada. Además, se pide a **A** cuantas otras personas que cumplen con las características de la encuesta conoce, si responde "30", se sabe que la probabilidad teórica de **B** de ser seleccionada por **A** fue de 1 en 30.

En el muestreo RDS, siguiendo el procedimiento para la recolección de datos a través de "olas" sucesivas o ciclos de reclutamiento, en un momento dado se alcanza el tamaño de la muestra que se denomina "**equilibrio**" con respecto a las variables que se miden. El **equilibrio** se puede interpretar, como un estado en el que convergen las estimaciones en torno a una estabilidad de la muestra, cuya composición no cambiará durante los siguientes ciclos de reclutamiento. En teoría, se alcanza el equilibrio dentro de las cuatro a seis olas de reclutamiento independientemente de las semillas iniciales. Una forma de ejemplificar esto se muestra con un estudio en jóvenes heterosexuales donde se seleccionan a 4 semillas varones, al cabo de 4 a 6 olas, la conformación de la muestra en la variable sexo es de 50% de mujeres y 50% de varones, alcanzando el **equilibrio** muestral.

Las poblaciones ocultas, como la población de este estudio o las trabajadoras sexuales, son fundamentales en una serie de problemas de salud pública. Sin embargo, debido a la naturaleza de estos grupos, es difícil obtener información precisa sobre ellos y esto complica los esfuerzos de prevención de enfermedades tales como el VIH/SIDA. El sistema de muestreo RDS mejora la capacidad para estudiar estas poblaciones ocultas, permitiendo hacer estimaciones de la prevalencia de ciertos rasgos en estas poblaciones.

2.2.2. Poblaciones ocultas

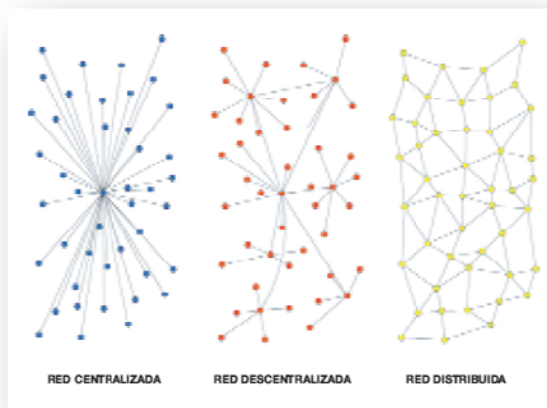
Las poblaciones ocultas tienen dos características básicas: primero, carecen de un marco muestral, por lo que su tamaño y sus márgenes reales son desconocidos; y segundo, las personas que pertenecen a esta población tienen un especial recelo a ofrecer información a los investigadores ya que normalmente siguen comportamientos estigmatizados, mal vistos o ilegales. Para la aplicación del RDS, la población objeto de estudio debe presentar tres requisitos⁸:

⁸ Heckathorn y Jeffri, 2005

- 1) Los informantes deben reconocerse los unos a los otros como miembros de la población objetivo pues, de lo contrario, no sabrían a quién seleccionar como nuevo informante;
- 2) Las redes sociales de los miembros de la población deben ser lo suficientemente densas como para garantizar una cierta profundidad sociométrica⁹. No obstante, esa "profundidad" tiene límites que, a su vez, generan restricciones inherentes a los métodos de muestreo realizados por encadenamiento, conseguidos a través de sucesivas oleadas de contactos y que, lógicamente, suelen relacionarse con el tamaño geográfico del área en el que se realiza el estudio; y
- 3) La población no debe estar muy segmentada en subgrupos, ya que las olas o encadenamientos que se generen a partir de los primeros informantes quedarían encapsuladas en los subgrupos.

2.2.3. Redes sociales

Figura 1. Redes Sociales



Una red social es un conjunto de nodos (personas, organizaciones y otras entidades sociales) que con las conexiones activas que los ligan (relaciones, confianza, entendimiento mutuo, valores compartidos, comportamientos e intercambio de información), posibilitan la acción cooperativa¹⁰ respecto a los individuos, adquieren una notable importancia, pues sirven de apoyo emocional, proporcionan contactos, son fuentes para la creación de alternativas, experimentación de nuevas ideas

y orientación o posibilidad de formación. En cierto modo son un punto intermedio entre las necesidades individuales y el entretejido social¹¹.

En su forma más simple, una red social es un mapa de todos los lazos relevantes entre todos los nodos estudiados. Se habla en este caso de redes "sociocéntricas" o "completas", en otras palabras una red **sociocéntrica** o completa se refiere a los patrones de relaciones dentro de un grupo definido, por ejemplo, una clase escolar, una congregación, etc.

Otra opción es identificar la red que envuelve a una persona (en los diferentes contextos sociales en los que interactúa); en este caso se habla de "red personal".

⁹ La sociometría es el estudio de la evolución de los grupos y de la posición que en ellos ocupan los individuos, prescindiendo del problema de la estructura interna de cada individuo / Levi Moreno, Jacob – *Fundamentos de Sociometría*, 1961.

¹⁰ Wasserman, 1994

¹¹ Zimmermann, 2004

Las redes personales representan todas las redes sociocéntricas a las que la persona pertenece (familia, trabajo, club deportivo, iglesia, etc.).

Los conceptos fundamentales de red social son los siguientes.

Actor: Son las entidades entre las cuales se establecen los vínculos que se pretenden analizar. Puede tratarse de individuos, empresas u otras unidades de carácter colectivo.

El nombre utilizado no implica que estas entidades necesariamente tengan la capacidad de resolución o de actuar.

Lazo relacional: Son los vínculos existentes entre pares de actores. La gama y tipo de lazos son muy diverso: opiniones de carácter personal (amistad, respeto, preferencia), transmisión de recursos (transacciones económicas, información), interacción entre individuos (hablar, escribirse), conexión física (una carretera, un puente), pertenencia o afiliación a una misma organización, relación de parentesco, etc.

Díada: Una díada consiste en un par de actores y los posibles vínculos entre ellos. Los vínculos se contemplan siempre como una propiedad de una pareja de actores, y nunca como una característica individual. Por lo tanto la díada es el nivel mínimo al cual puede realizarse el análisis en este trabajo.

Tríada: Subconjunto de tres actores y sus posibles vínculos. Importantes métodos y modelos se basan en ellas para su análisis, particularmente los interesados en la transitividad y en el equilibrio de las relaciones.

Subgrupo: Puede definirse como un subconjunto superior a tres actores y sus relaciones entre ellos.

Grupo: Sistema de actores que ha sido delimitado por razones conceptuales, teóricas o empíricas, lo cual permite ser tratado como un conjunto finito. Se trata del conjunto de actores cuyos vínculos serán analizados en este trabajo.

Red social: Conjunto finito de actores y de relaciones definidas entre ellos.

Algunos de los indicadores más utilizados en el análisis de redes sociales son¹²:

Indicador	Descripción
Grado (degree)	Cantidad de enlaces o conexiones directas de un nodo. Especificando la dirección se tiene grado de entrada (indegree) o lazos que llegan a un nodo y grado de salida (outdegree) o lazos que salen de un nodo.
Grado de intermediación (Betweenness)	Indica el número de veces que es necesario pasar por cada nodo para poder conectar otros dos, se hace contando los geodésicos (camino más cortos entre 2 nodos) existentes en la red y luego las veces que aparece cada nodo en ellos. Esta medida enfatiza el poder de los nodos para conectar recursos.

¹² **Indicadores de redes sociales.** A partir de Hanneman (2001), Molina (2002), Sunbelt (2001), Wasserman (1994).

Indicador	Descripción
Cercanía (Closeness)	Índice de la cercanía de un nodo con el resto de la red. Se calcula la suma de geodésicos que unen a cada vértice o nodo con el resto y se calcula su inversa.

Para el presente estudio se toma como indicador principal el Grado (Degree) para el cálculo de las probabilidades de transición que se describe más adelante.

2.2.3.1. Redes estructuradas de tipo centro periferia

En las redes de centro-periferia, normalmente hay unos pocos individuos muy populares, (los sociólogos se refieren a ellos como estrellas socio-métricas) en torno a las cuales muchas personas menos populares se reúnen. Un ejemplo de esto podría ser visto en la red de trabajadoras sexuales. La TS es la “estrella” o “centro” y sus clientes forman la periferia de la red. Del mismo modo, un agente (proxeneta) puede ser el núcleo de la red de TS’s con las mismas TS que forman la periferia.

Una red de centro-periferia se puede evaluar cuando la homofilia¹³ es positiva para los grupos de estatus más alto y negativa para los grupos de estatus inferior.

2.2.3.2. Redes estructuradas en cohorte

Las redes estructuradas en cohortes existen donde los miembros tienden a afiliarse principalmente con personas similares a ellos. Los miembros pueden clasificarse por edad, sexo o nivel educativo. El marcador para la evaluación de una estructura de cohortes es homofilia positiva para cada sub-grupo de interés.

2.2.3.3. Redes estructuradas bipartitas

Las estructuras bipartitas existen donde los individuos forman lazos con los diferentes a sí mismos, como las parejas sexuales de los heterosexuales. En las estructuras bipartitas, la homofilia será negativa para cada grupo.

2.2.4. Tipos de muestreo aplicables a estas poblaciones

2.2.4.1. Muestreo Bola de Nieve (snow-ball)

En las últimas dos o tres décadas, han surgido varios métodos para el reclutamiento de poblaciones ocultas para la vigilancia epidemiológica y otro tipo de investigaciones por encuestas.

¹³ La homofilia, para este estudio se utilizará para medir las *asociaciones de las redes sociales*, *homofilia perfecta*, en la que todos los lazos se forman dentro del grupo, se asigna el valor +1; *homofilia inexistente*, en la que los lazos se forman sin tener en cuenta la pertenencia al grupo, se asigna el valor cero, cuando todos los lazos se forman fuera del grupo, a la *homofilia* se le asigna el valor -1. Niveles intermedios de homofilia negativa son definidos de forma paralela a los niveles intermedios positivos.

Quizás el método más utilizado es el muestreo de bola de nieve, que implica la identificación de un número inicial de miembros del subgrupo de los cuales se tiene datos y que luego sirven como "semillas", para encuestar y/o para ayudar a identificar otros miembros del subgrupo (individuos que tienen el mismo tipo de conductas) que se incluirán en la muestra.

Estos individuos, a su vez se les pide que proporcionen información sobre los miembros de otro subgrupo y el proceso continúa hasta que el tamaño de muestra previsto se alcanza o en su defecto si la muestra se "satura" (es decir, miembros nuevos en la muestra del subgrupo no proporcionen más información que difiera de la obtenida de los miembros entrevistados anteriormente).

Aunque las semillas iniciales en el muestreo bola de nieve son elegidas al azar, en la práctica, esto es difícil sino imposible de realizar. Por lo tanto, en la práctica, las semillas iniciales en el muestreo bola de nieve tienden a ser elegidas a través de un muestreo por conveniencia. Al igual que otros métodos de muestreo no probabilístico, el principal inconveniente del muestreo de bola de nieve es el sesgo del muestreo, es decir, el peligro de que la muestra obtenida, no sea "representativa" de la población de la que se extrajo la muestra. En el muestreo de bola de nieve, la composición de la muestra está fuertemente influenciada por la elección de las semillas iniciales y este método también tiende a estar sesgado y favorecer la mayor cooperación de los entrevistados que forman parte de grandes redes personales, en contraposición a los sistemas de muestreo de elección probabilística.

Los métodos no probabilísticos de muestreo como el muestreo de bola de nieve, son útiles en la investigación formativa y en la definición del problema, pero no son aptos para la producción de datos que puede ser generalizado con confianza a una población mayor, aunque a veces (incorrectamente) se utiliza de esta manera.

2.2.4.2. Muestreo basado en establecimientos (Facility-Based)

El reclutamiento de miembros de una determinada población oculta en una variedad de entidades o servicios frecuentados por los miembros de estas poblaciones es otro método de uso común. Las correccionales han sido utilizadas para grupos de personas involucradas en actividades como el uso de drogas y el trabajo sexual.

Centros de salud que tratan ITS¹⁴ sirven a una gran proporción de HSH y trabajadoras sexuales, en Bolivia son los denominados CDVIR, encargados departamentales de la vigilancia epidemiológica para infecciones de transmisión sexual.

¹⁴ Infecciones de Transmisión Sexual

Cada una de estas entidades o servicios se han utilizado para reclutar a un gran número de miembros de las poblaciones ocultas, sin embargo, vienen con ciertos sesgos. Ninguno de los lugares proporcionan muestras probabilísticas que puedan considerarse representativas de una población dada. Las personas que tienen los medios para obtener los servicios particularmente serán diferentes a los miembros del grupo que asisten a estos servicios gratuitos.

2.2.4.3. Muestreo Dirigido

El muestreo dirigido, amplía las ideas del método bola de nieve, incluyendo una evaluación inicial por mapeo, que tiene como objetivo identificar diferentes redes o subgrupos que pueden existir en un determinado ambiente. Los subgrupos identificados de este modo son tratados como estratos de muestreo y cuando es posible se elige dentro de cada estrato a través de un muestreo sistemático. La magnitud del sesgo de muestreo en los muestreos dirigidos, dependen de la permeabilidad de la evaluación del mapeo. En la práctica, el tiempo y los recursos disponibles para llevar a cabo los mapeos limitan la utilidad de este método para la vigilancia epidemiológica.

2.2.4.4. Muestreo por lugar y tiempo (Time-Location Sampling - TLS)

Otro enfoque que ha tenido un uso creciente en los últimos años aprovecha el hecho de que algunas poblaciones ocultas tienden a reunirse o congregarse en determinados tipos de lugares o áreas geográficas, por ejemplo, las trabajadoras sexuales a menudo se congregan en lenocinios, salones de masajes y las esquinas de la calle, en la "luz roja"; los HSH se congregan en bares y "zonas" conocidas por atraer a HSH. En el muestreo TLS, estos sitios se enumeran en un mapeo etnográfico¹⁵ preliminar o en un ejercicio de evaluación pre vigilancia; la lista de lugares, se utiliza como un marco de muestreo, entre los que se elige una muestra probabilística de sitios y se realiza un segundo proceso de muestreo para la elección de las personas que serán encuestadas o en algún caso se realizará la entrevista a la totalidad de los miembros del subgrupo que se encuentran en el sitio determinado en un intervalo de tiempo preestablecido (por ejemplo, un período de tiempo escogido de forma aleatoria 3 horas en un día de la semana elegido también aleatoriamente).

Debido a que las probabilidades de selección se pueden calcular, TLS califica como un método de muestreo probabilístico.

Sin embargo, a menos que la totalidad o un gran porcentaje de los sitios donde se congregan los miembros del subgrupo sean identificados en el proceso de mapeo, para ser incluidos en el marco muestral y la totalidad o un gran porcentaje de los miembros del subgrupo visiten estos sitios

¹⁵ Se traduce como el análisis del modo de vida de un grupo de individuos, mediante observación y descripción de lo que la gente hace, como se comportan, cómo interactúan entre sí, para describir sus creencias, valores, motivaciones, perspectivas y como esto puede variar en distintos momentos y circunstancias, describiendo las múltiples formas de vida de los grupos involucrados.

periódicamente, este método sufre de niveles inaceptables de sesgo. Incluir todos los sitios en teoría se puede lograr, dando suficiente tiempo y recursos para desarrollo del marco muestral, pero también hay límites prácticos en cuanto a los recursos que se pueden comprometer en estas actividades de manera regular.

Debido a que los lugares donde los miembros de los subgrupos se congregan sufren cambios en el tiempo, es necesario repetir el ejercicio del desarrollo del marco muestral antes de cada recojo de información o ronda de vigilancia.

Disponer del marco de muestreo de las rondas de vigilancia anteriores reduce el costo del desarrollo de los marcos muestrales en las siguientes rondas, pero los costos de actualización del marco muestral tiende a ser bajo.

Sin embargo, el peligro real es que falten algunos o muchos sitios, lo que resultará en un potencial sesgo de muestreo, además un riesgo más serio representan los miembros del subgrupo que no visitan estos sitios, en este caso, ningún rigor en la construcción del marco muestral al reunir los sitios reducirá el sesgo y por lo tanto, si una proporción significativa de los miembros de un subgrupo determinado no suelen frecuentar estos sitios, el método TLS puede estar sujeto a un grave sesgo de muestreo.

Otra fuente importante de sesgo del TLS es la naturaleza de los sitios de reclutamiento, HSH que asisten a bares y discotecas que no desean participar en la encuesta porque no desean conocer su estado serológico¹⁶. Las trabajadoras sexuales que trabajan en una esquina de la calle no querrán perder un cliente potencial, lo anterior resume una actitud de rechazo a la encuestas por parte de la población indistinto de la precisión de la ubicación del sitio en el marco muestral y esto está asociado a un comportamiento particular de esta población y que es muy marcado, en poblaciones ocultas como la que se encuentra bajo estudio, la obligatoriedad de contestar a un cuestionario no es posible de aplicación, lo único que se conseguiría es un rechazo conjunto a realizar la encuesta y esto llevaría a la suspensión de todo el estudio.

2.2.4.5. Muestreo convencional por conglomerados

Cabe señalar que, en limitadas circunstancias, el muestreo por conglomerados convencional puede ser un método de muestreo adecuado para la vigilancia del VIH en las poblaciones en riesgo.

Para que el muestreo por conglomerados sea adecuado, es necesario disponer de un marco muestral relativamente completo de los miembros del grupo o tener la capacidad de construir uno. Por otra parte, es necesario poder acceder a todos los miembros del grupo durante el período de recopilación de datos. Estos requisitos pueden satisfacerse en el caso de las poblaciones fácilmente accesibles.

¹⁶ Presencia de antígenos o de anticuerpos específicos en un individuo, indicativo de la infección por un determinado microorganismo.

Otros grupos de interés potenciales (por ejemplo, transportistas, o grupos que ejerzan una actividad particular o una especialidad), a menos que se puedan hacer repetidas visitas, para poder realizar las entrevistas a miembros que no estén presentes en el momento de la recolección de datos podría dar lugar a un sesgo de no respuesta potencialmente grande, lo que hace el muestreo por conglomerados una opción imposible para grupos poblacionales ocultos que son de alto riesgo.

2.3. El trabajo con RDS

Como en otros métodos similares, la aplicación del RDS se inicia con la identificación de unos informantes iniciales clave que cumplen la función de “semillas”. Sin embargo, tres puntos lo hacen diferente¹⁷:

1. Las semillas no son seleccionadas aleatoriamente de la población objetivo, esto con el propósito principal de garantizar que puedan identificar a personas que cumplan con las características requeridas y estos puedan ser entrevistados, además de ser nuevos reclutados y que a su vez puedan reclutar a más personas pares. A las semillas se les solicita que seleccionen a personas de la población objetivo (o faciliten el contacto con el entrevistador), por lo que las semillas pueden formar parte o no de la población objetivo.

El proceso de selección es concebido como un proceso de Markov de primer orden de tal forma que las características de un nuevo informante dependen teóricamente de las características del informante que lo ha reclutado, pero no de las características de quien seleccionó al último reclutador. De este modo, la saturación de la muestra se obtiene cuando tras la sucesión de encadenamientos, se logra una estabilidad en la presencia porcentual de una serie de categorías grupales que se consideren significativas (sexo, grupos étnicos, grupos de edad, etc.).

Estas categorías deben ser mutuamente excluyentes, es decir, si se pertenece a la categoría de “afectados” por una determinada enfermedad no se puede formar parte al mismo tiempo de la de “no afectados”. Se asume que la estabilidad de la muestra, resulta independiente de la presencia relativa de las categorías de personas con las que se inició el proceso de reclutamiento¹⁸. En todo caso, a partir de la muestra real obtenida se creará matemáticamente una muestra teórica en equilibrio, que después se compara con la muestra real, para decidir si ésta es válida en términos estadísticos;

2. Debido a las reticencias derivadas de las peculiaridades de las poblaciones ocultas, algunos autores sugieren la puesta en práctica de un doble incentivo sobre los informantes: por participar como entrevistado (primary reward) y por reclutar a nuevos entrevistados (secondary reward). Los incentivos pueden ser de tipo monetario pero, a veces y en función de las características y necesidades de la población objetivo, se puede considerar la puesta en práctica de recompensas alternativas, en el caso de la presente investigación como recompensa se considera la atención médica que recibe el reclutado y no existe recompensa de tipo económico.

¹⁷ Heckathorn, 1997; Wang, Carlson, Falck, Siegal, Rahman y Li, 2005

¹⁸ Abdul-Quader et. al., 2006

3. Por medio de un sistema de cupones se limita la posibilidad a cada informante de seleccionar a más de tres futuros informantes, con el fin de anular los sesgos provocados por la presencia de reclutadores semi profesionales o por una voluntad de súper-colaboración de algunas personas, que provoque una sobre-representación de las redes de un individuo en concreto.

Se justifica este método dado que permite la identificación en cadena de redes sociales de gays, bisexuales y trans y otros hombres que tienen sexo con hombres para su incorporación a este estudio, logrando alcanzar a aquellas personas que no están vinculadas a las redes sociales tradicionales; sino aquellos hombres que se relacionan sexualmente con otros hombres a través de diferentes formas de comunicación entre ellos.

En general, cada encuestado se atribuye un cupón con un número de serie único, que le fue dado por otro encuestado. También tendrán un número limitado de cupones que pueden dar a otros. Así, es posible hacer un seguimiento de quien reclutó a quién. El hecho de que un entrevistado también es reclutador perpetúa la muestra en cadena.

El RDS requiere realizar un seguimiento del grado de cada encuestado. El grado de un nodo en una red es el número de conexiones a ese nodo, es decir, el número de vecinos de ese nodo. En el contexto del muestreo de referencia en cadena, el grado de un individuo se define como el número de personas que esta persona **podría** reclutar.

Se consideran sólo redes no direccionadas, de manera que el reclutamiento puede llevarse a cabo en ambas direcciones a través de una conexión de la red social. Las muestras de referencia en cadena son **con reemplazo**, es decir, cualquier individuo puede ser reclutado en la muestra más de una vez.

El muestreo con reemplazo en las muestras de referencia en cadena son diferentes del muestreo con reemplazo tradicional. La selección de cada unidad de muestreo está bajo el control de los propios encuestados, por lo tanto no son libres de volver a muestrear a un mismo encuestado en otro momento. En el presente estudio se asume que cada entrevistado es reclutado solo una vez, dado que la revisión médica no se realizará más de una vez a la misma persona, aunque los métodos teóricos del RDS están ideados para compensar el caso que se reclute más de una vez a la misma persona.

Además, suponemos que la fracción de muestreo es pequeña, de manera que podemos aplicar las soluciones para la toma de muestras con reemplazo. Además de los mencionados anteriormente, esta teoría se basará en los siguientes supuestos:

1. **Grado.** Los encuestados informan con precisión su grado en la red.
2. **El reclutamiento es al azar.** El reclutador selecciona a los nuevos reclutados de manera aleatoria uniforme de su red personal.
3. **Reciprocidad.** Las conexiones de red son recíprocas. Los encuestados reclutan a las personas con quienes tienen una relación preexistente, tales como conocidos, amigos o parejas sexuales regulares o no. Dichas conexiones son recíprocas, como ejemplo, *mis amigos y conocidos me consideran un amigo o conocido*.

En consecuencia, en términos de una red teórica, la red de reclutamiento potencial es no dirigida, por lo que si el encuestado **A** puede reclutar a **B**, entonces **B** también puede reclutar a **A**. Esto es requerido por el modelo de reciprocidad¹⁹ sobre la que el estimador RDS se basa. Esto se conoce formalmente como la *hipótesis de reciprocidad*.

4. **Convergencia.** El reclutamiento se modela como un proceso de Markov, donde el *estado* del proceso de Markov es el último individuo reclutado. Suponemos que el proceso de Markov es **irreductible** y que cada estado tiene un **tiempo de retorno finito**. Por lo tanto, existe un equilibrio único para el proceso de Markov y el reclutamiento converge rápidamente a este equilibrio. La implicación es que después de un pequeño número de pasos, la composición de la muestra pasa a ser independiente de los reclutadores iniciales ("semillas") que iniciaron el proceso de la referencia en cadena.

La condición **irreductibilidad** es equivalente a la condición de que la red social está bien comunicada, es decir, a cada nodo se puede llegar por un camino finito de cualquier otro nodo. Por otra parte, estas redes sociales se suponen finitas (aunque muy grandes), por lo que el tiempo de retorno esperado debe ser finito también.

A primera vista, la hipótesis de irreductibilidad puede parecer poco realista, sobre todo para poblaciones grandes, donde es muy probable que algunas unidades estén aisladas de la red en su conjunto. Se sabe por la teoría de redes que la mayoría de las redes poseen un denominado *componente gigante*, un grupo de nodos, tal que, existe una ruta en la red entre dos nodos distinta de 0 que tiende al infinito, tal como el tamaño de la población.

El *componente gigante* normalmente abarca a la gran mayoría de la población, siempre y cuando cumpla con algunas condiciones básicas. En gráficos aleatorios puros, el *componente gigante* constará del 99% de la población si los nodos tienen sólo 5 enlaces en promedio. Los estudios RDS superan normalmente este margen de holgura. Para los estudios con RDS se debe tener en cuenta que la inferencia estadística se limita al *componente gigante*, en lugar de la población total, si el componente gigante es muy grande (similar al tamaño poblacional) esto es una condición menor.

Además, la investigación en un problema de "small-world" o *mundo pequeño*²⁰ ha dado lugar a la observación de que casi todas las redes sociales tienen en promedio una longitud de trayectoria muy corta.

En consecuencia, hay relativamente pocos intermediarios entre dos individuos seleccionados al azar en la mayoría de las redes sociales. En redes aleatorias puras²¹, la longitud de la trayectoria crece logarítmicamente con el tamaño de la población. Es por lo tanto plausible que la probabilidad de selección de cualquier individuo en la red se establezca después de unos pocos reclutamientos, ya que a casi nadie en la población se puede llegar en un pequeño número de pasos.

¹⁹ Heckathorn 2002; Salganik y Heckathorn 2004

²⁰ Watts (1999)

²¹ Erdős y Renyi (1959); Newman (2001)

De hecho, es difícil o imposible hacer cumplir el reclutamiento aleatorio entre los encuestados y en muchos casos los encuestados pueden haber tenido razones especiales para la selección de un reclutado en particular. Sin embargo, el reclutamiento no aleatorio, si se produce, no necesariamente sesgara el estimador.

Mientras el reclutamiento no se correlacione con ninguna variable importante para la estimación (por ejemplo, el estudio de variables o grado), el efecto acumulado para el reclutamiento aparece como uniformemente aleatorio.

Un reclutamiento no aleatorio sería más obvio de evidenciar por la matriz de reclutamiento que sería sesgada y asimétrica. Sin embargo, es una fuente potencial de sesgo que se debe tener en cuenta.

2.3.1. Muestreo RDS como un proceso de Markov

Una muestra de referencia en cadena (*chain referral*), puede ser vista como un proceso estocástico en el que las características sociales de cada reclutador afectan a las características de los reclutados. En el caso de la preferencia sexual, significaría que los reclutadores de cada grupo sexual generarán una mezcla de grupos sexuales de reclutados.

El reclutamiento puede ser modelado como un proceso de Markov, una forma de proceso estocástico con dos características esenciales. La primera, el proceso puede asumir un número limitado de estados, por ejemplo, tres tipos de identidad sexual. La segunda, el proceso depende del estado, donde la probabilidad de pasar de un estado a otro depende de una matriz de probabilidades de transición, por ejemplo, la probabilidad de que el próximo recluta provenga de un determinado grupo, depende del grupo de los que el reclutador actual viene.

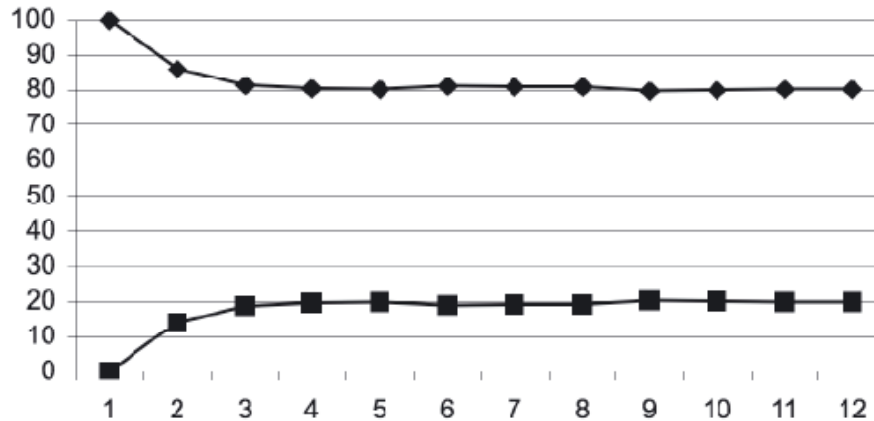
Por lo tanto este es un proceso sin memoria, en el que los patrones de reclutamiento sólo dependen del reclutador y no dependen del reclutado. Esto significa que el reclutamiento corresponde a un proceso de primer orden de Markov²². Asimismo, no hay grupos que sean reclutados exclusivamente desde el interior. Por lo tanto, el reclutamiento es "ergódico"; un proceso es denominado **ergódico** cuando es posible avanzar desde un estado a cualquier otro estado, no es necesario que esto sea en un solo paso. Cualquier estado puede repetirse y hay una probabilidad diferente de 0 de que cualquier estado vuelva a ocurrir.

Cuando se aplica a una muestra de *referencia en cadena*, los estados se refieren a las características de los sujetos, el paso de un estado a otro se refiere a un reclutador con un conjunto de características, reclutando a otro con características similares o diferentes; que cualquier estado puede repetirse significa que, después de una o más olas de reclutamiento, un reclutado puede tener las mismas características que el reclutador anterior. En esencia, esto significa que el reclutamiento no puede quedar atrapado dentro de un grupo o un conjunto de grupos, si ocurriese que la cadena de reclutamiento entra en este tipo de grupo no habría salida posible. Por lo tanto, los reclutamientos corresponden a un proceso "regular" de Markov.

²² Si un suceso depende de otro además del inmediatamente anterior, este es un proceso de Markov de mayor orden

Este modelo del proceso de reclutamiento, es relevante para entender la fiabilidad de los indicadores elaborados a partir del muestreo RDS a causa de dos deducciones, en relación con los procesos regulares de Markov. En primer lugar, la "ley de los grandes números de cadenas regulares de Markov" afirma que la probabilidad de que un sistema se encuentre en un estado determinado en el transcurso de un gran número de pasos es independiente de su estado inicial. La implicación para el reclutamiento RDS es:

Proposición Uno: *A medida que el proceso de reclutamiento continúa de ola en ola, eventualmente se alcanzará el equilibrio en la mezcla de los reclutados, independientemente de las características de los sujetos o conjunto de sujetos que comenzaron el reclutamiento.*



De acuerdo al gráfico se puede observar que el crecimiento de la muestra hace que ese alcance el equilibrio, indistinto de las semillas que sean seleccionadas, esto es evidente en este estudio y se muestra en la aplicación práctica, donde el crecimiento en la muestra se convierte en una distribución proporcional en los grupos bajo estudio.

Por lo tanto, si el reclutamiento opera hasta alcanzar un equilibrio y la red de reclutamiento resultante califica como un proceso regular de Markov, se evita el problema principal para el muestreo en poblaciones ocultas - que las características de la muestra reflejen la muestra inicial. En su lugar, una muestra RDS es totalmente independiente del conjunto inicial de sujetos.

En su lugar, el sesgo se plantea sólo cuando el muestreo no continúa a través de olas suficientes para alcanzar el equilibrio adecuado.

La restricción de que el proceso de reclutamiento debe pertenecer a un proceso de Markov de primer orden no es esencial. Por ejemplo, una muestra que no corresponda a un proceso de primer orden, sino que corresponda a un orden superior, no produce problemas graves.

El efecto que se produce sólo es una mayor lentitud en la aproximación al equilibrio, por lo que el muestreo puede requerir nuevas oleadas. Esta observación se puede convertir en limitativa sí la densidad de la población no es como se estimaba al inicio de la investigación, causando que no se puedan realizar más olas, en caso de ocurrir esto, no se podrá alcanzar el equilibrio y el tamaño de la muestra será insuficiente para realizar las estimaciones poblacionales.

La ley de los grandes números para cadenas regulares de Markov provee un medio para calcular el equilibrio analíticamente²³. El estado de **equilibrio** para un sistema con n tipos de sujetos ($E=E_a, E_b, \dots, E_n$), se encuentra resolviendo el sistema de n ecuaciones lineales.

$$\begin{aligned}
 1 &= E_a + E_b + \dots + E_n \\
 E_a &= S_{aa} E_a + S_{ba} E_b + \dots + S_{na} E_n \\
 E_b &= S_{ab} E_a + S_{bb} E_b + \dots + S_{nb} E_n \\
 &\vdots \\
 E_{n-1} &= S_{a,n-1} E_a + S_{b,n-1} E_b + \dots + S_{n,n-1} E_n
 \end{aligned} \tag{2.1}$$

Donde E_a, E_b, \dots, E_n son los valores proporcionales de equilibrio para los grupos a, b, ..., n, respectivamente y S_{xy} , es la probabilidad que un sujeto del tipo **x** reclute a un sujeto del tipo **y** o también llamada **proporción muestral**. Los estados de la primera ecuación que son las proporciones de equilibrio deben sumar 1. Las ecuaciones subsecuentes expresan los tamaños de los equilibrios de los grupos en función a los tamaños de equilibrio de los grupos y el reclutamiento de los grupos proporcionales de cada grupo. Debido a que este es un sistema de n ecuaciones con n incógnitas no tiene una solución única.

Por ejemplo para un sistema de 2 grupos es definido por el sistema de ecuaciones:

$$\begin{aligned}
 1 &= E_a + E_b \\
 E_a &= S_{aa} E_a + S_{ba} E_b
 \end{aligned}$$

Resolviendo se tiene

$$E_a = \frac{S_{ba}}{1 - S_{aa} + S_{ba}} \tag{2.2}$$

$$E_b = 1 - E_a \tag{2.3}$$

Un tema muy importante en la importancia práctica del teorema anterior es la velocidad a la que se alcanza el equilibrio, esta convergencia al equilibrio se producirá sin importar cuán lenta sea esta convergencia, sin embargo la velocidad de esta convergencia es bastante rápida, esta conclusión se basa en un teorema que demuestra que la convergencia se produce geoméricamente.

La implicación es:

²³ Kenelly y Snell (1960:72)

Proposición Dos: El conjunto de sujetos generado por una muestra RDS se aproxima al equilibrio a una tasa muy acelerada, geoméricamente.

Definición

Dada una sucesión a_n es posible formar una nueva sucesión S_n del siguiente modo:

$$\begin{aligned} S_1 &= a_1 \\ S_2 &= a_1 + a_2 \\ S_3 &= a_1 + a_2 + a_3 \\ S_4 &= a_1 + a_2 + a_3 + a_4 \\ &\dots \\ S_n &= a_1 + a_2 + a_3 + a_4 + \dots + a_n \end{aligned}$$

La sucesión S_n se llama serie y se denota por $\sum_{n=1}^{\infty} a_n$ o simplemente $\sum a_n$

Los elementos $a_1, a_2, a_3, \dots, a_n, \dots$ de la sucesión original son los términos de la serie y $S_1, S_2, S_3, \dots, S_n, \dots$ se denominan las sumas parciales de la serie. Una serie es una sucesión de sumas parciales.

Clasificación de una serie

- Si la sucesión S_n tiene límite finito S , la serie es convergente (converge a S).
- Si $\lim S_n = +\infty$ o $-\infty$ se dice que la serie es divergente.
- Si S_n no tiene límite, se dice que la serie es oscilante.

Serie geométrica

Aquella cuyos términos forman una progresión geométrica. (Cada término es igual al anterior multiplicado por una constante).

Se denomina a al primer término y k a la constante,

$$S_n = a + ak + ak^2 + ak^3 + \dots + ak^{n-1} = \sum ak^{n-1} \quad (2.4)$$

Se multiplica ambos miembros por k :

$$kS_n = ak + ak^2 + ak^3 + ak^4 + \dots + ak^n = \sum ak^n \quad (2.5)$$

Restando ambas ecuaciones:

$$S_n - kS_n = a - ak^n \quad (2.6)$$

$$S_n = \frac{(a - ak^n)}{(1 - k)} \quad (2.7)$$

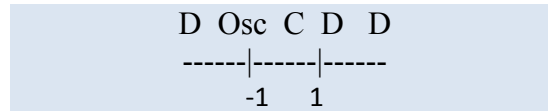
$$S_n = \frac{a}{1-k} - \frac{ak^n}{1-k} \quad (8)$$

Para $|k| < 1$ $\lim S_n = \frac{a}{(1-k)}$ pues $k^n \rightarrow 0$ la serie geométrica converge.

Para $|k| > 1$ la serie diverge pues $k^n \rightarrow \infty$.

Para $k = 1$ la serie diverge pues $S_n = na$.

Para $k = -1$ la serie es oscilante.



La implicación, es que la convergencia ocurre dentro de un pequeño número de olas de reclutamiento. Por lo tanto, la elección de una muestra diversa de semillas acelera la aproximación al equilibrio.

Así como la muestra se expande de ola en ola, la composición media de los sujetos en la ola principal, se aproxima al equilibrio a un ritmo geométrico. Sin embargo, como toda la muestra incluye tanto olas actuales como olas anteriores, la aproximación al equilibrio es menor a un crecimiento geométrico, aún así el crecimiento es bastante acelerado. El tamaño de este arreglo dependerá del número de reclutas por sujeto. Por ejemplo, si cada encuestado puede reclutar a 3 personas, el tamaño de la muestra se alcanzaría en 6 olas, sin embargo si el número de reclutados se eleva a siete, sólo son necesarias cinco olas.

Esto se debe a que cuanto mayor sea el número de reclutas por sujeto, mayor será la tasa a la cual el número de sujetos en cada ola se incrementa y por lo tanto menor será la influencia de las primeras olas en la composición final de la muestra.

Si el reclutamiento toma la forma de un proceso simple de Markov, es decir, una cadena lineal a partir de una sola semilla, la implicación es que las cadenas de reclutamiento serán largas, porque sólo entonces la mayoría de los sujetos serán extraídos en las olas, después de lo cual el equilibrio será logrado. Sin embargo, la producción de cadenas lineales (uno a uno), se enfrenta al problema que la mayoría de las cadenas no serán más largas que un máximo de tres pasos u olas debido al desgaste, cuando un encuestado no proporciona una referencia válida de otro posible reclutado.

Es por eso que en los métodos de referencia en cadena se permite a los encuestados dar referencias múltiples, lo que genera una red de reclutamiento en forma de árbol.

Este enfoque tiene dos ventajas. En primer lugar, ayuda a resolver el problema de la deserción, porque después de que la semilla ha producido varias referencias, el fracaso de un solo sujeto para producir referencias válidas no podrá detener el proceso de toma de muestras.

Por lo tanto, dando los recursos y tiempo suficiente, las cadenas de reclutamiento se pueden generar prácticamente de cualquier longitud. En segundo lugar, las referencias múltiples ayudan a reducir el sesgo de la elección de las semillas iniciales, porque, los reclutas pasan a ser socialmente más distantes de la semilla y también son más numerosos.

Hay una complicación potencial de la introducción de referencias múltiples, debido a que la estructura de reclutamiento en forma de árbol producida por las múltiples referencias no corresponde a una estructura lineal asumida por el modelo de cadenas de Markov. Para analizar formalmente las implicaciones de esta diferencia estructural para la aplicabilidad del modelo de Markov para los muestreos de referencia en cadena excedería el alcance de este trabajo, pero dos comentarios pueden ser hechos.

En primer lugar, una estructura en forma de árbol puede ser analizada como un conjunto de estructuras lineales, por ejemplo, un entrevistado de la ola final puede ser visto como el producto de una cadena lineal a partir de la semilla y cuyos vínculos corresponden a las olas intermedias. Parece razonable suponer que un análisis que es válido para una cadena lineal, será también válido para un conjunto de dichas cadenas.

En segundo lugar, se debe determinar empíricamente si el modelo de Markov se ajusta a los datos, mediante la comparación de la composición real de la muestra con la composición de la muestra teórica esperada, si el proceso de muestreo corresponde a un proceso de Markov, esto es, el **equilibrio (E)**, ecuaciones (2.2) y (2.3).

Puede parecer que las largas cadenas de referencia son innecesarias, dado que el equilibrio se puede calcular a partir de la matriz de probabilidades de transición y la matriz no depende de la longitud de las cadenas, pero sí del número de referencias para los reclutamientos.

Por lo tanto, se podría seguir la recomendación de Frank y Snijders (1994) y comenzar con un gran número de semillas seleccionadas por su diversidad y llevar a cabo una única ola. Sin embargo, este enfoque tiene dos desventajas.

En primer lugar, la muestra tendría una falta de profundidad sociométrica²⁴. La parte de la población oculta accesible puede no ser representativa de la población total. Incluso si se escoge por diversidad sexual, las semillas constituyen una muestra de conveniencia. Si solamente una sola ola se lleva a cabo, todos los sujetos se encuentran dentro de un solo enlace de los encuestados que son accesibles. Sectores socialmente más distantes de la población no aparecen en la muestra.

Por el contrario, cuando las cadenas de reclutamiento son de 12 o más pasos de largo, consistentemente con la literatura sobre el tema "*seis grados de separación*"²⁵ se debe poder acceder a todos los miembros de la población.

²⁴ La profundidad sociométrica puede ser definida como la cantidad de relaciones sociales que tienen las personas pertenecientes a un grupo que permite que el muestreo por referencia en cadena avance

²⁵ **Seis grados de separación** es una teoría que intenta probar el dicho de "el mundo es un pañuelo", dicho de otro modo, que cualquiera en la Tierra puede estar conectado a cualquier otra persona del planeta a través

En segundo lugar, las cadenas de referencia largas son eficientes, porque los encuestados que no son semilla y que no son miembros de la ola final desempeñan un doble papel, ya que a la vez son la fuente y el producto de la referencia.

Cuanto más larga sea la cadena de referencia, mayor es el número de encuestados intermedios y por lo tanto mayor es la proporción de referencias a los encuestados.

En RDS, cadenas de referencia largas se producen de dos maneras. Primero, a los encuestados se les premia por el reclutamiento. Segundo, se imponen cuotas de reclutamiento máximo para que ningún pequeño subconjunto de reclutadores pueda monopolizar los derechos de reclutamiento: la cuota en el presente estudio se fijó en tres reclutas. Las cuotas de reclutamiento se llevaron a cabo utilizando un sistema de cupones, donde los reclutadores potenciales dan cupones a sus reclutas. El cupón incluye un número de serie que documenta la relación entre el reclutador y el recluta. La limitación a tres cupones para el reclutamiento garantiza que todos los reclutas están cerca del inicio de la cola de reclutas potenciales.

La introducción de incentivos al reclutamiento no sólo alarga las cadenas de referencia, sino que también reduce el sesgo debido al exceso de voluntarizaje²⁶. Además de los incentivos al reclutamiento, se debe aprovechar la presión del grupo al motivar a los compañeros del entrevistado potencial y así emplear la influencia social.

En esencia, los incentivos para el reclutamiento sirven como transformadores que convierten los estímulos materiales en incentivos simbólicos basados en los pares (es decir, la influencia social ejercida por los reclutadores).

En suma, las muestras de referencia en cadena pueden producir indicadores fiables, esto sucede cuando la muestra alcanza su equilibrio independientemente del punto de inicio. Lo que se requiere es que las cadenas de reclutamiento sean lo suficientemente largas para aproximarse al equilibrio. En el método RDS, las cadenas de referencia se producen por un mecanismo de reclutamiento denominado "*Participante Dirigiendo el Reclutamiento*" que combina incentivos al reclutamiento y las cuotas. Sin embargo, hay ciertas limitaciones para evaluar la variabilidad de los indicadores cuantitativos. Sin embargo existen medios para calcular los errores estándar de las estimaciones de población.

2.3.1.1. RDS, un proceso de cadenas de Markov mediante métodos de Monte Carlo (MCMC)

Detrás de este método de MCMC está una cadena de Markov en un espacio de estados V .

En el contexto de RDS, V es la población que será muestreada y en la aplicación actual, se limita al caso en que V es una población finita de tamaño N . Se identifica la cadena como $K(v_i, v_j)$ que es la probabilidad de transición desde el estado v_i al estado v_j , donde:

de una cadena de conocidos que no tiene más de cinco intermediarios (conectando a ambas personas con sólo seis enlaces).

²⁶ Heckathorn (1997)

$$K(v_i, v_j) \geq 0, \quad \sum_{v_j \in V} K(v_i, v_j) = 1 \quad (2.9)$$

En términos de RDS, $K(v_i, v_j)$, es la probabilidad que un individuo v_i reclute un individuo v_j . La cadena es irreductible (es decir, existe una única clase de estados) si para cada par de puntos v_i, v_j , hay una probabilidad de que partiendo de v_i se llegue a v_j . Bajo este supuesto, existe una distribución única $\pi: V \rightarrow \mathbb{R}$ denominada **distribución estacionaria**, que satisface:

$$\sum_{v_j \in V} \pi(v_i) K(v_i, v_j) = \pi(v_j) \quad (2.10)$$

Esto es, si X_0, X_1, X_2, \dots es una realización de la cadena con $X_0 \sim \pi$, entonces $X_i \sim \pi$ para $i \geq 0$. En consecuencia, a partir de una cadena en equilibrio, la caminata puede ser usada para generar muestras dependientes de la misma distribución π .

Una muestra de referencia en cadena puede ser usada para obtener muestras dependientes de la población V con distribución π :

$$P(X_i = v_j) = \pi(v_j) \quad (2.11)$$

Que quiere decir que para cada individuo muestreado v_j tiene una probabilidad $\pi(v_j)$ de ser elegido. Entonces para cualquier función $f: V \rightarrow \mathbb{R}$ la media muestral es:

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \quad (2.12)$$

Que es un estimador insesgado, no de la media muestral, pero sí de $\mathbb{E}_\pi f = \sum_{i=1}^N f(v_i) \pi(v_i)$. Esto es porque las unidades son seleccionadas con probabilidades desiguales, la media muestral no es un estimador consistente de la media poblacional. Como es común en la literatura de muestreo, la idea detrás de la importancia del muestreo es que la media muestral ponderada

$$\frac{1}{n} \sum_{i=1}^{n-1} \frac{f(X_i)}{N \cdot \pi(X_i)} \quad (2.13)$$

produzca un estimador insesgado de la media poblacional μ_f de f donde

$$\mathbb{E}_\pi \left(\frac{f(X_i)}{N \cdot \pi(X_i)} \right) = \sum_{i=1}^N \frac{f(v_i)}{N \cdot \pi(v_i)} \pi(v_i) \quad (2.14)$$

$$= \frac{1}{N} \sum_{i=1}^N f(v_i) \quad (2.15)$$

En particular, si $D \subseteq V$ es un subconjunto de individuos que cumplen con la característica, entonces (2.12) puede ser utilizado para estimar la proporción, $p = \frac{|D|}{N} = \mu_f$ por consiguiente $f(v_i) = 1$ si $v_i \in D$ y $f(v_i) = 0$ en otro caso.

A menudo es necesario reemplazar (2.12) por el estimador asintóticamente insesgado.

$$\hat{\mu}_{AI} = \frac{1}{\sum_{i=0}^{n-1} \frac{1}{\pi(X_i)}} \sum_{i=0}^{n-1} \frac{f(X_i)}{\pi(X_i)} \quad (2.16)$$

La considerable ventaja de (2.16) sobre (2.12) es que la importancia del peso $\frac{1}{\pi(X_i)}$ solo necesita ser evaluada como una constante, por lo que no es necesario conocer el N poblacional. En la aplicación de RDS esto es esencial.

2.3.1.2. Aplicación a RDS

La importancia del muestreo permite la estimación de p , dadas las muestras X_0, X_1, X_2, \dots a partir de una distribución fija π . RDS genera las muestras a través de un proceso similar a MCMC.

Se considera una red social $G=(V,E)$ donde los nodos $x \in V$ representando a individuos en la población que por ejemplo pueden ser infectados o sanos, y $e \in E$ que representa los lazos en la red. Se asume lazos simétricos ponderados (relacionamiento simétrico) y se escribe $W(x, y) = W(y, x)$ para el peso del lazo entre los nodos x e y .

Además, se asume que la red está conectada (es decir, que existe un camino entre cada par de individuos de la población).

Para un subconjunto de individuos $A \subseteq V$ se usa la siguiente notación para denotar el peso de A .

$$W_A = \sum_{x \in A} \sum_{y \in V} W(x, y) \quad (2.17)$$

Se modela el procedimiento de muestreo RDS como caminata aleatoria en el grafo G definido por $K(x, y) = \frac{W(x, y)}{W_x}$ donde $K(x, y)$ es la probabilidad que un individuo x reclute a un individuo y ²⁷. Se asume que la red está

²⁷ Se debe tener en cuenta el RDS en la práctica se lleva a cabo con un muestreo sin reemplazo (es decir, los que participan no pueden participar de nuevo), aunque se define como un muestreo con reemplazo.

conectada y la cadena es irreducible y el camino tiene una única distribución estacionaria²⁸.

$$\pi(x) = \frac{W_x}{W_V} \quad (2.18)$$

Consecuentemente, para X_0, X_1, X_2, \dots la realización de la cadena con $X_0 \sim \pi$ y $f: V \rightarrow \mathbb{R}$ ²⁹ cualquier función, el estimador de muestreo (2.16) de la media de la población μ_f se reduce a

$$\hat{\mu}_f = \frac{1}{\sum_{i=0}^{n-1} \frac{1}{W_{X_i}}} \sum_{i=0}^{n-1} \frac{f(X_i)}{W_{X_i}} \quad (2.19)$$

En el caso de la estimación de la prevalencia de una enfermedad, se tiene $f(v_i) = 1$ si v_i está infectado y $f(v_i) = 0$ en otro caso. Simplificando (2.19)

$$\hat{p} = \frac{1}{\sum_{i=0}^{n-1} \frac{1}{W_{X_i} X_i \text{ infectado}}} \sum \frac{1}{W_{X_i}} \quad (2.20)$$

Para evaluar los estimadores de RDS (2.19) y (2.20) todavía se necesita conocer los pesos W_{X_i} . Típicamente se propone un peso uniforme para el lazo $W(x, y) = 1$, correspondiente a la hipótesis de que los participantes reclutan a sus contactos de manera uniformemente al azar y que todos los contactos están de acuerdo con participar. **En este caso, W_x es igual al grado del nodo**³⁰.

Las estimaciones muestrales ponderadas de RDS proporcionales a su probabilidad de selección. En el caso de que todos los nodos tengan el mismo grado, la media muestral estimada será equivalente a la estimación RDS (2.19), dado el supuesto del reclutamiento uniforme.

Para lo anterior, se parte de que la caminata es estacionaria: $X_0 \sim \pi$, es decir, la semilla inicial se extrae de acuerdo a una distribución estacionaria. Sin embargo, si la caminata es aperiódica, es decir, si la red es no bipartita, entonces el estimador de RDS $\hat{\mu}$ es asintóticamente insesgado, independientemente de la distribución de partida. Por otra parte, hay un teorema central de límite para $\hat{\mu}$.

$$\sqrt{n}(\hat{\mu}_n - \mu) \rightarrow N(0, \sigma_f^2)$$

²⁸ Se dice que una Cadena de Markov en tiempo discreto admite una distribución estacionaria en la medida que las probabilidades de largo plazo existen y son **independientes de la distribución inicial (f_0)**.

²⁹ Aunque en los estudios RDS a menudo se utilizan semillas múltiples, por simplicidad se supone que una sola semilla es seleccionada. Sin embargo, la mayoría de los resultados se pueden generalizar para manejar múltiples semillas sin complicaciones significativas y se mantienen sin cambios cualitativos.

³⁰ Conocer grado de un individuo es en sí mismo es un reto, en particular en el contexto de RDS, y constituye una fuente de error no muestral.

Para cualquier distribución inicial en X_0 . El teorema de límite central para cadenas de Markov es aplicado típicamente a estimadores insesgados (2.12), sin embargo se puede generalizar este resultado para aplicar a (2.19), primero se debe observar que:

$$\begin{aligned} \sqrt{n}(\hat{\mu}_n - \mu) &= \frac{\sqrt{n}}{\sum_{i=0}^{n-1} \frac{1}{W_{X_i}}} \sum_{i=0}^{n-1} \frac{f(X_i) - \mu}{W_{X_i}} \\ &= \left[\frac{1}{\frac{1}{n} \sum_{i=0}^{n-1} \frac{W_V}{NW_{X_i}}} \right] \left[\frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} \frac{f(X_i) - \mu}{\frac{NW_{X_i}}{W_V}} \right] \end{aligned} \quad (2.21)$$

Por la ley de los grandes números de cadenas de Markov, el primer término en la expresión final converge a 1. El resultado sigue siendo aplicable al teorema de límite central de Markov al segundo término con la función:

$$g(X_i) = \frac{f(X_i) - \mu}{\frac{NW_{X_i}}{W_V}} \quad (2.22)$$

La varianza σ_f^2 depende de la varianza de f y la estructura de autocorrelación de la cadena puede ser dificultosa en la práctica.

Cabe señalar que estos resultados sobre el comportamiento asintótico de las estimaciones de RDS dependen fundamentalmente de la validez de los supuestos del modelo. En particular, estos resultados requieren que los participantes recluten a una persona elegida al azar de manera uniforme de su red de contactos y que los participantes puedan ser reclutados en tiempos de muestra múltiple (es decir, muestreo con reemplazo).

Además, incluso si todas las condiciones adecuadas se cumplen, la teoría asintótica dice poco sobre el rendimiento del RDS en pequeñas muestras ($n \approx 200$). En el caso de muestras pequeñas, la estructura de red social es de vital importancia y está asociada a su profundidad sociométrica.

2.3.2. Modelo de homofilia, evaluación del sesgo

La fiabilidad es una condición necesaria, pero no condición suficiente para un indicador eficaz, un indicador fiable puede estar sesgado. Se analiza un tipo de sesgo sistemático en los datos de las muestras de referencia en cadena, debido a la **homofilia (h)** o endogamia³¹.

³¹ La endogamia sirve para medir las asociaciones de las redes sociales, a mayor endogamia, mayor nivel de asociatividad y por lo tanto la concentración en la muestra es mayor. Se entiende como endogamia el rechazo a la incorporación de miembros ajenos a un grupo social en particular.

La idea esencial de la teoría, es que los *vínculos sociales* de un sistema social estructurado serán no aleatorios: algunas relaciones serán más probables que otras, donde los "sesgos" se refieren a cualquier desviación de un patrón completamente aleatorio de la conexión.

Estos sesgos pueden consistir en una tendencia hacia la afiliación en grupo, como suele ocurrir en las amistades o una tendencia a la afiliación a grupos, como en los sistemas de *relaciones exogámicas*³². Este hecho considera las probabilidades de selección desiguales.

En un grupo no estructurado, el reclutamiento sólo reflejará la prevalencia de cada sub grupo dentro de la población oculta. Cada reclutador, es independiente de la identidad del grupo que recluta.

En un sistema donde la afiliación de grupo afecta a la selección, los miembros reclutados por cada individuo reflejarán los prejuicios tanto del reclutador y la prevalencia de diferentes tipos de miembros dentro de la población. La selección de tal sistema puede ser modelado como un proceso con dos pasos condicionales³³.

Así conceptualizada, la *magnitud de la homofilia* reflejará una mezcla de factores culturales y situacionales, que van desde el énfasis cultural en la afiliación dentro del grupo de pertenencia, a la facilidad de transporte entre grupos geográficamente distintos.

También puede verse afectada por la dirección de los incentivos, como los utilizados para aumentar el reclutamiento de ciertos grupos específicos.

A partir de Fararo y Sunshine (1964), Blau (1977, 1994), y los modelos de Rapaport (1979), la homofilia fue definida formalmente como: *Homofilia perfecta*, en la que todos los lazos se forman dentro del grupo, se asigna el valor +1; *Homofilia inexistente*, en la que los lazos se forman sin tener en cuenta la pertenencia al grupo, se asigna el valor cero. Como ejemplo, si los lazos que se forman dentro del grupo son equivalentes a un tercio de las veces y los lazos que se forman fuera del grupo son dos tercios, el nivel de *homofilia* es 0.33.

El concepto de homofilia puede ser ampliado para cubrir el caso donde exista un sesgo por la formación de lazos intragrupos, a esto se denomina *heterofilia*, cuando todos los lazos se forman fuera del grupo, a la *homofilia* se le asigna el valor -1. Niveles intermedios de homofilia negativa son definidos de forma paralela a los niveles intermedios positivos.

A fin de especificar el modelo de homofilia, consideremos el caso de un sistema de dos grupos, **A** y **B**, donde la homofilia es positiva. La probabilidad de que un miembro del grupo **A** seleccione dentro de su grupo es, S_{aa} , la probabilidad de que la selección sea controlada por la homofilia, H_a , y la probabilidad de que la

³² La regla establece que la pareja obligatoriamente debe ser elegida entre los miembros de un grupo ajeno al propio, *prohíbe* las relaciones intra-grupo. La situación inversa se denomina endogamia donde la regla determina el grupo dentro del cual se debe implementar la relación, es decir, *impone* un grupo específico. Rodríguez García, Dan (2002)

³³ Fararo y Skvoretz (1984:233)

homofilia no ocurra es $(1 - H_a)$, ponderados por el porcentaje de miembros del grupo **A** en la población, P_a , es decir:

$$S_{aa} = H_a + (1 - H_a)P_a \quad (2.23)$$

Por principios similares, la probabilidad de que un miembro del grupo **A** seleccione a un miembro del grupo **B**, es la probabilidad de que la homofilia no ocurra que es $(1 - H_a)$, ponderado por la proporción de los miembros del grupo **B** en la población, P_b , es decir,

$$S_{ab} = (1 - H_a)P_b \quad (2.24)$$

$$S_{ba} = (1 - H_b)P_a \quad (2.25)$$

Extender este modelo a los casos donde la homofilia es negativa requiere pequeñas modificaciones en estas expresiones. Por ejemplo, cuando un grupo es heterófilo, requiere de la conjunción de dos eventos, que la heterofilia no determine la formación del lazo, un evento con una probabilidad de $1 + H_d$, y que un miembro del grupo sea seleccionado independientemente de la identidad de grupo, un evento cuya probabilidad depende del tamaño proporcional del grupo, P_a , la probabilidad de formar un lazo intra grupo es el producto de las probabilidades de estos dos eventos, es decir,

Si $H_a < 0$, entonces

$$S_{aa} = (1 + H_a)P_a \quad (2.26)$$

El valor absoluto de la homofilia $|H_a|$, es la probabilidad de que la homofilia direcciona la formación de lazos. Cuando la homofilia es positiva, es la probabilidad que un lazo se forme intra grupo. Cuando la homofilia es negativa, es la probabilidad de que un lazo se forme desde fuera del grupo.

El cálculo de la homofilia está dado por

$$H_a = \frac{D_a - (S_{ba}D_b) - (S_{ab}D_a)}{D_a} \left. \vphantom{H_a} \right\} \text{Si } S_{aa} \geq P_a \quad (2.27)$$

$$H_a = \frac{D_a - (S_{ba}D_b) - (S_{ab}D_a)}{S_{ba}D_aS_{ab}} \left. \vphantom{H_a} \right\} \text{Si } S_{aa} < P_a \quad (2.28)$$

$$H_b = \frac{S_{ba} - P_a}{-P_a} \quad (2.29)$$

El objetivo es identificar el modelo robusto³⁴ más simple. El modelo Fararo-Skvoretz capta las dos características esenciales de la homofilia, a) que los grupos varían en su grado de homofilia y b) que las selecciones resultantes producen la estructura de afiliaciones intra grupo y extra grupo dentro del sistema. Con este conjunto de ecuaciones, el modelo especifica la relación entre la **proporción de la población** oculta (P) y la **probabilidad de selección** (S).

Basándose en la combinación de este modelo de homofilia y el modelo de Markov, se deriva un teorema que demuestra la **distribución de equilibrio (E)**, igual a la **distribución de la proporción poblacional (P)**, si la **homofilia (H)** es igual. En ese caso, el equilibrio proporciona una estimación de la población insesgada. En esencia, lo que muestra es que los grupos con alta homofilia son sobre muestreados, pero el sobre muestreo se anula, si todos los grupos tienen igual homofilia. El equilibrio converge con la distribución de la población y por lo tanto (E) se convierte en un estimador insesgado de (P).

Basado en las ecuaciones (2.23) y (2.24), el equilibrio de la muestra se puede expresar como una función tanto de la población y las condiciones de homofilia. Para simplificar el análisis se considera en primer lugar el caso de dos subgrupos. De acuerdo con la ley de los grandes números para cadenas regulares de Markov, la proporción de equilibrio de miembros de cada grupo se encuentra resolviendo un sistema de dos ecuaciones lineales mencionadas en la proposición 1:

$$I = E_a + E_b \quad (2.30)$$

$$E_a = S_{aa} E_a + S_{ba} E_b \quad (2.31)$$

Resolviendo se tiene

$$E_a = \frac{S_{ba}}{1 - S_{aa} + S_{ba}} \quad (2.32)$$

Esto a su vez puede ser ampliado por la sustitución de las ecuaciones (2.23) y (2.24) para expresar el equilibrio (E) de la muestra, en términos de homofilia (H) y población (P):

$$E_a = \frac{P_a(1 - H_b)}{1 - (H_a + P_a(1 - H_a)) + P_a(1 - H_b)} \quad (2.33)$$

Simplificando se tiene:

³⁴ Un modelo robusto es aquel que no se ve afectado por variaciones pequeñas respecto a la hipótesis, esto quiere decir que no son afectados indebidamente por valores atípicos u otras pequeñas discrepancias respecto de las asunciones del modelo

$$E_a = \frac{P_a(H_b - 1)}{P_a H_b - P_a H_a + H_a - 1} \quad (2.34)$$

Esta expresión proporciona una base para derivar conclusiones basadas tanto en la teoría de redes parciales y las cadenas de Markov, con respecto a las condiciones en que el RDS, producirá una muestra insesgada, es decir, **una muestra en la que la proporción de cada grupo en la muestra es igual a la proporción del grupo en la población.** Es decir.

Proposición Tres: *Un sistema RDS extrae muestras insesgadas, si todos los términos de homofilia de todos los grupos son iguales, es decir, para cualquier grupo X, $E_x = P_x$, si y sólo si $H_x = H_y$ para cualquier otro grupo Y.*

Esto se demuestra a través de:

Teniendo en cuenta que la homofilia afecta al reclutamiento, la composición del equilibrio del grupo (E) no tiene necesariamente que corresponder a la verdadera distribución de la población oculta (P).

La medida en que los miembros de cualquier grupo serán muestreados depende de tres factores: a) el tamaño del grupo, b) su tendencia a la endogamia y c) la fuerza de la endogamia en los otros grupos. La derivada parcial³⁵ muestra que la proporción de un grupo en la muestra aumenta y todo lo demás es constante a medida que aumenta el tamaño del grupo y la homofilia.

Del mismo modo, disminuye la proporción de un grupo con el incremento de la endogamia u homofilia de otros grupos. Por lo tanto, un pequeño grupo potencialmente podrá parecer grande en la muestra RDS si su homofilia es fuerte y si los miembros de otros grupos tienen homofilia débil. Por el contrario, un grupo grande puede parecer pequeño en la muestra RDS si su homofilia es débil y si otros grupos tienen fuerte homofilia. La efectividad del RDS como un medio para elaborar muestras insesgadas depende de si estas dos posibilidades son plausibles.

Las condiciones en las que el RDS produce muestras imparciales se pueden deducir a través del análisis de la ecuación (2.34). Esto puede hacerse mediante la identificación de las condiciones en que E_a es igual a P_a . En primer lugar, se substituye P_a por E_a , en la ecuación (2.34) anterior:

³⁵ La derivada parcial proporciona un medio para determinar si la relación entre dos términos es consistentemente positiva, siempre negativo, o mixta. El primer paso es calcular la primera derivada de un término con respecto al otro. La dirección de la relación se halla entonces por determinar si, dadas las limitaciones en los valores de los parámetros, la derivada es siempre positiva, siempre negativa o mixta. Por ejemplo, para determinar la relación entre el tamaño de la muestra de equilibrio y el sesgo de la homofilia del grupo, derivar E, respecto de H, es decir:

$$dE_a/dH_a = \frac{P_a(1 - H_b)(1 - P_a)}{(P_a H_b + P_a H_a + H_a - 1)^2}$$

$$P_a = \frac{P_a(H_b - 1)}{P_a H_b - P_a H_a + H_a - 1} \quad (2.35)$$

Al resolver para H_a

$$H_a = \frac{H_b(P_a - 1)}{P_a - 1} \quad (2.36)$$

Simplificando

$$H_a = H_b \quad (2.37)$$

Esta conclusión significa que cuando la homofilia es igual, los efectos *inflacionarios* de la homofilia de cada grupo son compensados exactamente por los efectos *deflacionarios* de la homofilia de los otros grupos.

Así, los dos términos de homofilia se cancelan mutuamente, dejando el tamaño de la muestra determinado exclusivamente por el tamaño de la población.

Aunque esta prueba se aplica al caso de dos grupos se puede extender al caso de tres grupos y ha sido confirmado por simulaciones en sistemas más grandes.

Una limitación de esta proposición es notable, cuando los términos de homofilia son muy grandes, que reflejan el aislamiento mutuo de los grupos del sistema, la aproximación al equilibrio es más lenta, significa que se deberá hacer más olas en el reclutamiento para lograr la aproximación al equilibrio. Por lo tanto, teniendo en cuenta las limitaciones prácticas que restringen el número de olas de reclutamiento, el equilibrio se alcanzará sólo cuando la homofilia no sea extrema. La implicación es que cuando los límites que separan los grupos son visualmente inalterables, RDS debe ser utilizado para tomar muestras dentro de esos grupos y no entre ellos, incluso los términos de homofilia se debe probar que son iguales.

La igualdad en términos de homofilia es una condición muy estricta. Por lo tanto, la pregunta de cómo los términos de homofilia están relacionados unos con otros es crucial para evaluar la capacidad del RDS para evitar sesgos. Desafortunadamente, la teoría de redes parciales no proporciona una guía respecto a las relaciones en términos de homofilia.

Los argumentos teóricos sugieren que la homofilia alta en un grupo tiende a producir homofilia alta en otros grupos, lo que reduce las diferencias de homofilia, es decir que cuando un solo grupo refuerza sus límites, induce a otros grupos a hacer lo mismo³⁶.

³⁶ Simmel (1955)

Así, en un sistema en el que otros contribuyen a fortalecer su grado de homofilia, otro grupo reaccionará aumentando su propia homofilia o endogamia. Por el contrario, en un sistema donde la homofilia sea pequeña, el grupo tenderá a evitar la homofilia. El efecto acumulado de estos procesos crea una relación positiva entre los términos de homofilia. Si esta teoría es correcta, la hipótesis de igualdad de la homofilia es similar en muchos sistemas sociales.

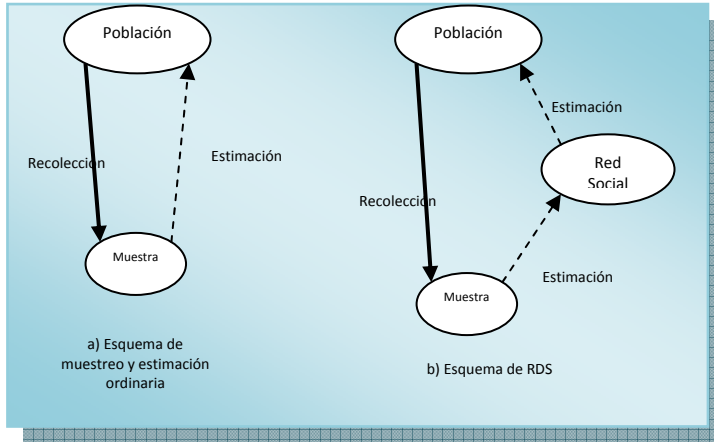
Hay que señalar que la presentación inicial de RDS había una fuerza no reconocida, ofreciendo medios para controlar el sesgo debido al reclutamiento diferencial. Este sesgo se produce cuando un grupo de reclutas especialmente eficientes y su distintivo patrón de reclutamiento son excesivamente representados en la muestra, aunque las cuotas de reclutamiento reducen esta forma de sesgo potencial, puesto que no todos los sujetos cumplen sus cuotas, la variación se mantiene.

Esta fuente de sesgo no afecta el equilibrio del modelo de Markov, porque no depende del número absoluto de reclutas de cada grupo, sino más bien de la distribución proporcional de los reclutas (es decir, los términos S). Estas son las que impulsan el modelo, porque las probabilidades de transición se basan en la distribución proporcional de los reclutas de cada grupo, las probabilidades siguen siendo las mismas si todos los grupos reclutan por igual o algunos grupos reclutan más o menos que otros.

2.4. Estimaciones poblacionales

2.4.1. Uso de las redes sociales para realizar estimaciones hacia la población

Figura 2. Proceso de estimación RDS



El procedimiento de estimación no se hace directamente a la población, la muestra se utiliza para hacer estimaciones sobre la red que conecta a la población. Luego, utilizando la información sobre esta red, se deriva la proporción de población en distintos grupos. Al no intentar estimar directamente de la muestra a la población, se evitan problemas de las muestras de referencia en cadena. Para estimar de la red social a la población, se utilizará el *modelo de reciprocidad* que se introdujo en Heckathorn

(2002). El modelo de reciprocidad proporciona una estimación del tamaño proporcional de la población oculta basado en dos fuentes de datos: las probabilidades de transición derivadas del análisis de los patrones de reclutamiento y el reporte de tamaño de la red personal. Cuando la estimación poblacional es derivada desde el modelo de reciprocidad se compara con el equilibrio o distribución muestral, los resultados a partir de este modelo son consistentes con que los grupos de redes más grandes tendrán mayor proporción en la muestra.

El modelo de reciprocidad en un sistema con N grupos puede ser representado por un sistema de ecuaciones, en el que los primeros estados en que los tamaños proporcionales de población deben sumar uno y los otros expresan el principio de reciprocidad para cada uno de los pares de grupos, donde el número de pares es $(N(N - 1)) / 2$. Por lo tanto, un sistema con cuatro grupos es descrito por siete ecuaciones, como sigue:

$$\begin{aligned}
 1 &= P_a + P_b + P_c + P_d \\
 P_a \cdot D_a \cdot S_{ab} &= P_b \cdot D_b \cdot S_{ba} \\
 P_a \cdot D_a \cdot S_{ac} &= P_c \cdot D_c \cdot S_{ca} \\
 P_a \cdot D_a \cdot S_{ad} &= P_d \cdot D_d \cdot S_{da} \\
 P_b \cdot D_b \cdot S_{bc} &= P_c \cdot D_c \cdot S_{cb} \\
 P_b \cdot D_b \cdot S_{bd} &= P_d \cdot D_d \cdot S_{db} \\
 P_c \cdot D_c \cdot S_{cd} &= P_d \cdot D_d \cdot S_{dc}
 \end{aligned}
 \tag{2.38}$$

Aquí los términos **D** se refieren a los tamaños de la red, los términos **S** derivan de la matriz de transición, y los términos de **P** son las estimaciones de población que se deriva del modelo de reciprocidad. Un sistema de ecuaciones lineales deben tener el mismo número de ecuaciones e incógnitas, sin embargo, aquí hay siete ecuaciones y sólo cuatro incógnitas (es decir, los términos P), por lo que el sistema está sobredeterminado.

Si el ajuste entre el modelo de reciprocidad y los datos fueran perfectos, bastaría simplemente elegir arbitrariamente cuatro ecuaciones y resolver los cuatro términos de P. Sin embargo, dado que el ajuste con los datos reales nunca es perfecto, la elección de las ecuaciones es lo que importa.

El método estándar para resolver sistemas sobredeterminados es lineal por mínimos cuadrados³⁷, un procedimiento que emplea la misma lógica que la regresión lineal. La ventaja del enfoque de mínimos cuadrados lineales es que se basa en un método estadístico estándar para resolver los conflictos entre las ecuaciones.

La teoría del muestreo de cadenas de Markov se ha desarrollado con el fin de muestrear a partir de distribuciones arbitrarias. La premisa es diseñar un proceso de Markov de tal manera que la distribución de equilibrio del proceso de Markov sea idéntica a la distribución de la que se desea hacer la muestra. Se ha demostrado que los estimadores basados en muestras de cadenas de Markov son asintóticamente insesgados (sección 2.3.1.1 de este documento)³⁸.

En contraste con el muestreo tradicional de cadenas de Markov, no se está en libertad de diseñar las probabilidades de transición entre las unidades de muestreo debido a la falta de un marco de muestreo estándar.

Más bien las probabilidades de transición se imponen por la naturaleza de la muestra de referencia en cadena y las propiedades de la red social. Sin embargo, la muestra de referencia en cadena constituirá una cadena de Markov que se ajusta a los criterios que rigen la aplicación de la teoría.

En términos matemáticos, una muestra de referencia en cadena es análoga a una caminata aleatoria³⁹ en una red.

Se ha demostrado⁴⁰ que una caminata aleatoria en una red es un proceso de Markov, que en equilibrio ocupa un nodo con probabilidad proporcional al grado.

³⁷ Farebrother 1988

³⁸ Por asintóticamente insesgado se entiende que cualquier sesgo será del orden de $1/n$. Por lo tanto, para tamaños de muestra significativa, cualquier sesgo será despreciable (Hastings 1970).

³⁹ En su forma más general, las caminatas aleatorias son cualquier proceso aleatorio donde la posición de una partícula en cierto instante depende sólo de su posición en algún instante previo y alguna variable aleatoria que determina su subsecuente dirección y la longitud de paso.

$X(t)$ define una trayectoria que empieza en la posición $X(0)=X_0$. Una caminata aleatoria se modela mediante la siguiente expresión:

$$X(t + \tau) = X(t) + \Phi(\tau)$$

donde Φ es la variable aleatoria que describe la ley de probabilidad para tomar el siguiente paso y τ es el intervalo de tiempo entre pasos subsecuentes. A medida que la longitud y dirección de un paso dado depende solo de la posición $X(t)$ y no de alguna posición previa, se dice que la caminata aleatoria posee la Propiedad de Márkov.

Se puede entonces inferir que una muestra de referencia en cadena seleccionará individuos de la población con probabilidad proporcional al grado.

Se considera una población hipotética, la población se compone de dos grupos de personas (por ejemplo, VIH + y VIH-), se estimará la proporción de la población en cada uno de estos grupos. Se puede observar que el número de vínculos de alguien en el grupo **A** a alguien en el grupo **B**, que en este ejemplo es 6, es el mismo como el número de vínculos de alguien de **B** a alguien en el grupo **A**. Esta afirmación puede parecer trivial, pero resulta ser muy útil porque hay dos maneras diferentes de calcular este número de lazos entre grupos.

Sea X la matriz de adyacencia de la red, X tendrá elementos x_{ij} donde $x_{ij} = 1$ si los nodos i y j están conectados y será igual a cero en otro caso. Se consideran sólo las relaciones recíprocas, por lo que sí es el caso de que $x_{ij} = 1$, entonces es también el caso de que $x_{ji} = 1$. Se llaman a estas, *relaciones de amistad*, aunque hay otras relaciones posibles. Se debe tener en cuenta que el **grado del nodo i** , d_i , es la suma de la i -ésima fila de X ,

$$d_i = \sum_j x_{ij} \quad (2.39)$$

Se define el **grado de una persona i** , d_i , como el número de amistades de la persona i .

Si el camino aleatorio es en el nodo i en t pasos, la probabilidad de que el nodo i elija el nodo j es

$$1/d_i = 1/\sum_j x_{ij} \quad (2.40)$$

Se denota esta probabilidad de transición como $K(v_i, v_j)$ y sea la matriz con estas probabilidades de transición llamada σ^X . La caminata al azar en la red puede considerarse un proceso de Markov con probabilidades de transición, $K(v_i, v_j)$ es la probabilidad de que cualquier individuo v_i reclute al individuo v_j .

El número total de las amistades que se irradia de las personas en el grupo **A**, R_A , es la suma del grado de todas las personas en el grupo **A** y se define como

$$R_A = \sum_{i \in A} d_i = N_A \cdot D_A \quad (2.41)$$

Donde N_A es el **número de personas** en el grupo A y D_A es el **grado medio** de personas en el grupo A .

⁴⁰ Salganik y Heckathorn (2004)

Los caminos aleatorios que se consideran son irreducibles⁴¹ y finitos, por lo que debe haber un equilibrio único del proceso de Markov. Además, el proceso de Markov convergerá hacia el equilibrio. Se considera el vector de estado x^* con elementos

$$x_i^* = \frac{d_i}{\sum_j d_j} \quad (2.42)$$

Se puede verificar que x^* es un equilibrio para el proceso de Markov dado por σ^X , y por la hipótesis, también debe ser un único equilibrio de atracción. Ahora que se ha establecido que una muestra de referencia en cadena del tipo RDS es una muestra de cadenas de Markov, se puede proceder a desarrollar estimadores para la población objetivo.

Usando sólo el hecho de que las muestras RDS individuales con probabilidad proporcional al grado, se puede desarrollar un estimador de tipo Hansen-Hurwitz⁴² (HH) para **P, el porcentaje de la población de cada conjunto disjunto A, B, ..., N.** La derivación que se presenta aquí utiliza un argumento similar al presentado en Salganik, Heckathorn (2004), para estimar el grado medio en una red social a partir de datos de referencia en cadena.

Los estimadores HH requieren el conocimiento de las probabilidades de selección, **S_i , la probabilidad de que un individuo i se seleccione** en cualquier etapa de la muestra de referencia en cadena. Los estimadores HH también asumen que cada elemento de la muestra se elige de forma independiente del resto de la muestra, un supuesto que es violado por el modelo de proceso de Markov en el reclutamiento. Así, la semejanza con los estimadores HH es sólo parcial. Las unidades de la muestra estarán correlacionadas, un hecho que no será el sesgo del estimador, sino que se deriva una estimación de la varianza.

Usando la condición de equilibrio, ecuación (2.42), las **probabilidades de selección** serán

$$S_i = \frac{d_i}{NDu} \quad (2.43)$$

Se puede estimar como

⁴¹ Un subconjunto cerrado $C \subseteq S$ se dice que es irreducible si y solo si no contiene ningún subconjunto propio cerrado

⁴² Estimación de Hansen-Hurwitz es un procedimiento estándar cuando se estima a partir de datos donde el muestreo es con reemplazo y las unidades cuentan con probabilidades desiguales de ser seleccionadas (Hansen y Hurwitz, 1943; Cochran, 1977; Brewer y Hanif, 1983). La idea básica, similar al estimador de Horvitz-Thompson, es que el procedimiento ponderado a cada elemento de la muestra por la inversa de su probabilidad de selección. Es decir, las unidades con pequeña probabilidad de ser seleccionadas cuentan más.

$$\hat{s}_i = \frac{d_i}{N\widehat{D}_u} \quad (2.44)$$

Donde \widehat{D}_u es el estimador de la media del grado del total de la población.

Los \hat{d}_i son fáciles de estimar. El grado medio puede ser estimado como un estimador de razón de estimadores HH.

$$\widehat{D}_u = \frac{\sum_L d_i/ns_i}{\sum_L 1/ns_i} = \frac{n}{\sum_L \frac{1}{d_i}} \quad (2.45)$$

Y sólo para un grupo, por ejemplo; un subconjunto **A** dentro de la población.

$$\widehat{D}_A = \frac{n_A}{\sum_{i=1} \frac{1}{d_i}} \quad (2.46)$$

Esta es la fórmula de la media armónica⁴³, la media de una cantidad con probabilidad proporcional a su tamaño.

La variable y_i , con algunos de los verdaderos valores de la variable de interés. Sea **T_y** que representa **el valor total de y en la población** $\sum_i y_i$ donde y_i puede representar variables continuas o variables dicotómicas tales como es el estatus de VIH.

El estimador HH del total y en la población, \hat{T}_y es

$$\hat{T}_y = \frac{1}{n} \sum_L \frac{y_i}{\hat{s}_i} = \frac{1}{n} \sum_{i \in L} \frac{\widehat{D}_u N y_i}{d_i} = \frac{\widehat{D}_u N}{n} \sum_L \frac{y_i}{d_i} \quad (2.47)$$

Si N es desconocido, como es generalmente el caso, todavía se puede estimar el valor medio de **y** como

$$\hat{y} = \frac{\widehat{D}_u}{n} \sum_L \frac{y_i}{d_i} \quad (2.48)$$

Substituyendo la definición de \widehat{D}_u , ecuación (2.45), se llega a la ecuación

⁴³ La media armónica, denominada H, de una cantidad finita de números es igual al recíproco, o inverso, de la media aritmética de los recíprocos de dichos números

$$\hat{y} = \frac{\sum_{i \in L} \frac{y_i}{d_i}}{\sum_{i \in L} \frac{1}{d_i}} \quad (2.49)$$

En esencia la ecuación (2.49) pondera cada caso por el recíproco del valor del grado correspondiente.

Interesa estimar P_A , la proporción de la población de tipo A. Sea y_i la función indicadora $I_A(i)$, la cual toma el valor 1 si $i \in A$, 0 en otro caso.

Usando la ecuación (2.49) se tiene

$$\hat{P}_A = \frac{\sum_{i \in A \cap L} \frac{1}{d_i}}{\sum_{i \in L} \frac{1}{d_i}} \quad (2.50)$$

Hay una forma alternativa de la ecuación (2.50), que da una idea de cómo funciona el estimador. Se tiene

$$\hat{P}_A = \left(\frac{n_A}{n} \right) \cdot \left(\frac{\widehat{D}_u}{\widehat{D}_A} \right) \quad (2.51)$$

La primer parte de la ecuación (2.51) (n_A/n) , es el porcentaje de la muestra de tipo A. Si la muestra fuera aleatoria normal esto sería la estimación de P_A . La segunda parte, $(\widehat{D}_u/\widehat{D}_A)$, es la corrección debida al efecto de la red. Por ejemplo, si $\widehat{D}_u > \widehat{D}_A$ se tienen individuos sub muestreados de tipo A y consecuentemente se inflará el estimador.

Se debe tener en cuenta que los reclutas iniciales en una muestra de referencia en cadena, las "semillas", serán elegidas de manera no aleatoria. Por lo general es prudente excluirlas del estimador, ecuación (2.50), así como la estimación de grado medio, ecuación (2.45), aunque el estimador asintóticamente insesgado se dará incluso si están incluidas. Cualquier sesgo potencial que pueda derivarse de la selección inicial de las semillas se reduce. Los reclutamientos realizados por las semillas se incluyen en la matriz de reclutamiento.

Se considera para una determinada red X , la probabilidad de seguir una *amistad* escogida de forma aleatoria a partir de una persona en el grupo A que termine en alguien del grupo B , se puede definir como S_{AB} , se tiene

$$S_{AB} = \frac{T_{AB}}{R_A} \quad (2.52)$$

Donde T_{AB} es el **número de vínculos** que tiene una persona en el grupo **A** y una persona en el grupo **B**.

Puesto que se considera sólo los lazos de reciprocidad, se sabe que el número de relaciones del grupo **A** al grupo **B**, es el mismo que el número de relaciones del grupo **B** al grupo **A**. Se puede calcular este número de dos maneras diferentes:

- 1) El **número de amistades que irradian** desde el grupo **A**, R_A , la probabilidad de que eventualmente una de esas relaciones lleven a alguien en el grupo **B**, S_{AB} o
- 2) El **número de amistades que irradian** desde el grupo **B**, R_B , la probabilidad de que eventualmente una de esas relaciones involucre a alguien del grupo **A**, S_{BA} . Es decir,

$$R_A S_{AB} = T_{AB} \quad (2.53)$$

$$R_B S_{BA} = T_{AB} \quad (2.54)$$

Se ajusta las ecuaciones (2.53) y (2.54) iguales entre sí y con la definición de la R_A y R_B , podemos escribir

$$N_A D_A S_{AB} = N_B D_B S_{BA} \quad (2.55)$$

Se debe tener en cuenta que la ecuación (2.55) reúne a información sobre las características de los nodos y las características de la red.

Sin embargo, incluso si se tuviera la información completa acerca de la red social - es decir, si se conociera D_A , D_B , S_{AB} , y S_{BA} - todavía se tendría una ecuación con dos incógnitas N_A y N_B , el tamaño de la población en el grupo **A** y grupo **B**. Se divide ambos lados de la ecuación (2.55) por N , el tamaño total de la población, entonces se reescribe la ecuación (2.55) en términos de proporciones de la población, P_A y P_B . Esto permite añadir una segunda limitación - que la suma de las proporciones de la población debe ser 1. Así se tiene

$$P_A \cdot D_A \cdot S_{AB} = P_B \cdot D_B \cdot S_{BA} \quad (2.56)$$

$$P_A + P_B = 1 \quad (2.57)$$

Donde P_A es la **proporción de la población en el grupo A** y P_B es la **proporción de la población en el grupo B**.

Ahora se tiene un sistema con dos ecuaciones y dos incógnitas. Usando álgebra ordinaria, se deduce que

$$P_a = \frac{D_a \cdot S_{ba}}{D_a \cdot S_{ab} + D_b \cdot S_{ba}} \quad (2.58)$$

$$P_b = \frac{D_a \cdot S_{ab}}{D_a \cdot S_{ab} + D_b \cdot S_{ba}} \quad (2.59)$$

Al examinar las ecuaciones (2.58) y (2.59) se revela que se puede recuperar las proporciones de la población, P_a y P_b , conociendo solo la estructura de la red que conecta a la población.

Se puede ver que es posible estimar la proporción de la población en el grupo **A** y en el grupo **B**, pero sólo si se conoce algo de información sobre la red de conexión de las personas en estos grupos. A continuación se analiza los métodos para recoger una muestra que se puede utilizar para estimar la información de la red.

2.4.2. Uso de la muestra para realizar estimaciones acerca de las redes sociales

Una vez seleccionada la muestra, se debe tener un procedimiento para usar la información de dicha muestra para hacer estimaciones acerca de la red social de la cual se ha extraído. El trabajo previo en la estimación de propiedades de las redes a menudo requiere una muestra aleatoria de nodos⁴⁴.

Sin embargo, no es posible recoger una muestra de una población oculta. Esta imposibilidad obliga a hacer estimaciones a partir de una muestra de referencia en cadena.

2.4.2.1. Supuestos

Con el fin de hacer estimaciones a partir de una muestra RDS, primero se deben hacer algunas suposiciones acerca de la población objeto de estudio y la forma en que el reclutamiento se realiza, haciendo estos supuestos explícitos. Es útil pensar en la selección de la muestra como un proceso que alterna entre la selección de nodos y la selección de vértices.

Es decir, los primeros nodos son extraídos para formar la ola 0 de la muestra. A continuación, estos nodos eligen los vértices que definen el período 1 de reclutamiento. Los vértices reclutados en el periodo 1 determinan los nodos extraídos en la ola 1. El proceso continúa de esta manera con los nodos de la selección de vértices que a su vez seleccionan los nodos hasta que se alcanza el tamaño de muestra deseado.

Para hacer más clara la hipótesis, primero se debe desarrollar una misma notación. Se comienza por la creación de dos funciones indicadoras: una para los nodos, $NI(j)_{w=x}$, que indica si un nodo dado, j , se selecciona en la ola X y otra para la dirección de los vértices, $EI(e_j \rightarrow k)_{r=x}$ que indica la dirección del vértice dado $e_j \rightarrow k$ se selecciona en el periodo de reclutamiento X . Estas funciones indicadoras se definen como sigue:

⁴⁴ Frank 1981

$$NI(j)_{w=x} = \begin{cases} 1, & \text{si el nodo } j \text{ es seleccionado en la ola } x \\ 0, & \text{en otro caso} \end{cases} \quad (2.60)$$

$$EI(e_j \rightarrow k)_{r=x} = \begin{cases} 1, & \text{si el vértice } e_j \rightarrow k \text{ es seleccionado en el reclutamiento } x \\ 0, & \text{en otro caso} \end{cases} \quad (2.61)$$

En primer lugar, se considera solamente el caso del muestreo con reposición, a pesar de que en la práctica real de la toma de muestras a veces es sin reemplazo. Este supuesto simplifica los cálculos, ya que elimina cambios en la población como en el progreso de la muestra.

Se asume que la red de la población oculta forma un componente conectado. Es decir, que hay un camino entre una persona y otra. Esto puede parecer como un supuesto restrictivo, pero una serie de resultados de la teoría de grafos aleatorios predice que incluso para gráficos muy esparcidos, casi todos los nodos pertenecen a un componente gigante⁴⁵. Para muchas poblaciones ocultas, las redes de amistad utilizadas en el reclutamiento son lo suficientemente densas para que este supuesto sea razonable.

Es importante señalar que la hipótesis de que la red forma un sólo componente, es la única hipótesis acerca de la estructura de la red. Esta falta de potenciales supuestos no comprobables de la red, hace que el muestreo RDS sea aplicable a muchos tipos diferentes de poblaciones ocultas.

Además, se supone que todos los entrevistados reciben y utilizan un cupón y que cuando los demás encuestados reclutan a otros, ellos reclutan aleatoriamente de todos los vértices que los involucran. Para especificar este supuesto de reclutamiento con mayor claridad, se tiene:

$$Pr[EI(e_j \rightarrow k)_{r=x+1} = 1 \mid NI(j)_{w=x} = 1] = \frac{1}{d_j} \quad (2.62)$$

Por último, se supone, en principio, que las semillas se extraen con una probabilidad proporcional a su grado. Es decir, una persona con 10 amigos tiene el doble de probabilidades de ser una semilla contra una persona con 5 amigos. Este supuesto se puede expresar como

$$Pr[NI(j)_{w=0} = 1] = \frac{d_j}{\sum_{i \in N} d_i} \quad (2.63)$$

Se opta por este supuesto ya que en estudios de poblaciones ocultas, la gente que se extrae como semillas son a menudo aquellos que son conocidos por los investigadores. Estas personas más conocidas tienden a tener más amigos que la media. Por lo tanto, parece razonable suponer que las probabilidades de una persona de ser seleccionada como semilla incrementan con su grado.

⁴⁵ Ver página 14 de este documento

2.4.2.2. Consecuencias de estos supuestos

Como se asume que las semillas se seleccionan con probabilidad proporcional al grado, se puede hacer algunas conclusiones sobre la probabilidad de que ciertos vértices serán seleccionados en un periodo de reclutamiento 1 y entonces la probabilidad de que ciertos nodos serán extraídos en la ola 1.

Una relación $e_{j \rightarrow k}$ puede ser seleccionada en el período de reclutamiento 1 sólo si el nodo desde el que se señala, nodo j , es seleccionado en la ola 0. Utilizando la notación de la función de indicadores y las reglas de probabilidad condicional, se puede calcular la probabilidad de que una determinada relación, $e_{j \rightarrow k}$, será seleccionada en el período de reclutamiento 1 como sigue:

$$Pr[EI(e_{j \rightarrow k})_{r=1} = 1] = Pr[NI(j)_{w=0} = 1] Pr[EI(e_{j \rightarrow k})_{r=1} = 1 | NI(j)_{w=0} = 1] \quad (2.64)$$

Como se ha supuesto que las semillas fueron seleccionadas con probabilidad proporcional a su grado y que las personas eligen al azar de sus relaciones al hacer un reclutamiento, podemos reescribir la ecuación (2.64) como

$$Pr[EI(e_{j \rightarrow k})_{r=1} = 1] = \frac{d_j}{\sum_{i \in N} d_i} \cdot \frac{1}{d_j} \quad (2.65)$$

Qué se puede simplificar a,

$$Pr[EI(e_{j \rightarrow k})_{r=1} = 1] = \frac{1}{\sum_{i \in N} d_i} \quad (2.66)$$

La ecuación (2.66) dice que si los nodos de la ola 0 se seleccionan con probabilidad proporcional al grado, a continuación, cada relación tiene la misma probabilidad de ser seleccionada en un período de reclutamiento.

Por lo tanto, la probabilidad de que un nodo, j , se seleccionará en la ola 1 es igual a la suma de las probabilidades de que las relaciones, d_j , que conduce a él se seleccionará en un período de reclutamiento uno. Es decir,

$$Pr[NI(j)_{w=1} = 1] = \sum_{d_j} \frac{1}{\sum_{i \in N} d_i} = \frac{d_j}{\sum_{i \in N} d_i} \quad (2.67)$$

La ecuación (2.63) muestra que si las semillas se seleccionan con probabilidad proporcional al grado, entonces, los nodos en la ola 1 serán seleccionados también con probabilidad proporcional al grado.

La repetición de este argumento de forma iterativa muestra que, si los nodos de la ola 0 se seleccionan con probabilidad proporcional al grado, a continuación, los nodos en todas las sucesivas oleadas se seleccionaran también con probabilidad proporcional al grado.

Este argumento también se puede repetir iterativamente para demostrar que, si las semillas se seleccionan con probabilidad proporcional al grado, entonces la probabilidad de que un vértice específico, $e_{j \rightarrow k}$, será seleccionado en el periodo de reclutamiento X constante e igual para todos los vértices:

$$Pr[EI(i)_{r=x} = 1] = \frac{1}{\sum_{i \in N} d_i} \quad (2.68)$$

Es importante señalar que estos argumentos se aplican a cualquier estructura de la red con los nexos de reciprocidad y por lo tanto son muy generales.

2.4.3. Construyendo los estimadores

Con los resultados obtenidos en la sección anterior, se derivan estimadores para las propiedades de la red específica y se demuestra que estas estimaciones son asintóticamente insesgadas a través de las cadenas de Markov y más adelante a través de estimadores Hansen-Hurtwitz.

Primero, se muestra cómo utilizar el comportamiento del reclutamiento observado, para estimar la **probabilidad de conexiones cruzadas** entre grupos, S_{AB} y S_{BA} . En segundo lugar, se estima el **grado medio** para las personas en los grupos, D_A y D_B , utilizando la información reportada del tamaño de la red. En tercer lugar, se combinará estas estimaciones para estimar la **proporción de la población** perteneciente a uno de dos distintos grupos, P_A y P_B .

2.4.3.1. Estimación de las probabilidades de selección recíprocas (S_{AB} y S_{BA})

Se debe recordar que se quiere utilizar la información de la muestra para estimar algunas propiedades de la red conectada a la población. Lo primero que se busca estimar es la probabilidad que si se sigue una amistad aleatoria iniciada en el grupo A y se atraviesa los grupos esta termine en el grupo B . Una forma de estimar esta probabilidad, S_{AB} , sería pedir a los encuestados los porcentajes de amigos que pertenecen a ciertos grupos. Sin embargo, esto no es posible porque para muchos temas de interés - por ejemplo el estado serológico de VIH - los entrevistados pueden no tener suficiente información acerca de sus compañeros para hacer evaluaciones precisas.

En lugar de basar la probabilidad de amistad entre grupos en datos auto-reportados, las basamos en la conducta real. Cuando uno de los encuestados recluta a otro, este vínculo conductual representa un eslabón en la red que se pueden verificar pidiendo a los reclutados la relación con el reclutador para la caracterización. La verificación requiere que el reclutado identifique al reclutador como un conocido, amigo o algo más que amigo y no como un extraño. Sólo los enlaces verificados deben ser utilizados para la estimación.

Para ser capaces de construir una estimación de S_{AB} y S_{BA} , debemos saber algo acerca de cómo el reclutamiento observado ha sido seleccionado a partir del conjunto de los posibles reclutamientos. Cada vértice, $e_{j \rightarrow k}$ tiene la misma probabilidad de ser seleccionado en cada período de reclutamiento. Es decir, los reclutamientos que observamos son una muestra aleatoria de todos los posibles reclutamientos. Recuerde que para cada muestra observamos un conjunto de reclutamientos.

Estos reclutamientos pueden ser divididos en cuatro grupos: reclutamientos de una persona en el grupo **A** a otra persona en el grupo **A**, r_{AA} , reclutamientos de una persona en el grupo **A** a una persona en el grupo **B**, r_{AB} , etc.

Dado que los reclutamientos observados son una muestra aleatoria de todos los vértices, las estimaciones insesgadas para S_{AB} y S_{BA} son:

$$\hat{S}_{AB} = \frac{r_{AB}}{r_{AA} + r_{AB}} \quad (2.69)$$

$$\hat{S}_{BA} = \frac{r_{BA}}{r_{BB} + r_{BA}} \quad (2.70)$$

2.4.3.2. Estimación del grado medio del grupo (D_A y D_B)

Durante el proceso de toma de muestras se recoge el grado de cada miembro de la muestra. Si se trata de estimar el **grado medio** de personas en el grupo **A** se tiene el grado medio de las personas en la muestra, la estimación será muy elevada, en algunos casos demasiado alta porque los métodos de referencia en cadena sobre representan a las personas con un grado alto⁴⁶.

Debido a que la media simple no es un buen estimador, es necesario un procedimiento de estimación diferente. Es decir, se debe tomar los datos de la muestra y ajustarlos, de manera que produzcan información precisa sobre la población. Dos enfoques distintos se pueden utilizar para motivar la exactitud de este ajuste, y, como se verá, estos dos enfoques llevan al mismo estimador. **Entonces quedará demostrado que este estimador es asintóticamente insesgado.**

Un enfoque para la construcción de un estimador para el grado medio, es el enfoque de distribución del grado y está dado por el grado de distribución de la muestra y la población. Si los supuestos se cumplen, entonces los nodos son seleccionados con probabilidad proporcional a su grado en todas las olas. Se utiliza este hecho, junto con la distribución de la muestra del grado observado, $q_A(d)$, para estimar la distribución del grado poblacional $p_A(d)$.

Esta distribución del grado poblacional puede ser usada para estimar el grado medio de personas en el grupo **A**, D_A .

⁴⁶ Erickson 1979; Kalton y Anderson 1986; Eland-Goosensen et al. 1997

Se ha encontrado que si los nodos se seleccionan con probabilidad proporcional a su grado, la distribución del grado muestral, $q_A(d)$, está dada por⁴⁷;

$$q_A(d) = \frac{d \cdot p_A(d)}{\sum_{d=1}^{\max(d)} d \cdot p_A(d)} \quad (2.71)$$

Donde $p_A(d)$ es la distribución del grado poblacional y

$$\sum_{d=1}^{\max(d)} d \cdot p_A(d) \quad (2.72)$$

Es una constante normalizada para asegurar que la suma de $q_A(d)$ sea 1

La ecuación (2.71), permite predecir la distribución del grado muestral dada la distribución del grado poblacional. Sin embargo, se tiene el problema opuesto. Es decir, se conoce la distribución del grado de la muestra y se desea predecir la distribución del grado de la población.

Dado que, como se describe en la ecuación (2.71), $d \cdot p_A(d)$ es proporcional a $q_A(d)$, también es el caso de que $p_A(d)$ es proporcional a $\frac{1}{d} \cdot q_A(d)$. Por lo tanto, si una muestra tiene una distribución del grado $q_A(d)$, entonces la distribución del grado poblacional, $P_A(d)$, puede estimarse como:

$$\widehat{p_A(d)} = \frac{\frac{1}{d} \cdot q_A(d)}{\sum_{d=1}^{\max(d)} \frac{1}{d} \cdot q_A(d)} \quad (2.73)$$

Donde $\sum_{d=1}^{\max(d)} \frac{1}{d} \cdot q_A(d)$ es una constante normalizada.

A partir de la estimación de la distribución poblacional, $\widehat{p_A(d)}$, se puede estimar la media del grado en la población, D_A , recordando que la media de una función de densidad de una probabilidad discreta $f(x)$ es $\sum_{x=0}^{\infty} x \cdot f(x)$.

Dado que se usa una aproximación de la distribución poblacional, se anota el estimador, $\widehat{D_A}$ con el super índice *dist*

$$D_A^{dist} = \sum_{d=1}^{\max(d)} d \cdot \widehat{p_A(d)} \quad (2.74)$$

⁴⁷ Feld 1991; Newman, Strogatz y Watts, 2001; Newman 2003a

En la ecuación (2.74) la sumatoria en el estimador esta indexada por el grado y no por el elemento de la muestra. Si se vuelve a escribir la ecuación (2.74) indexada por los elementos de la muestra, se obtiene:

$$\widehat{D}_A^{dist} = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}} \quad (2.75)$$

El estimador presentado en la ecuación (2.75) a primera vista parece diferente de un estimador de la media poblacional debido a que el tamaño de la muestra, n_A , está en el numerador y no en el denominador.

Sin embargo, este estimador es equivalente a la razón de dos estimadores de Hansen-Hurwitz⁴⁸, una que estima R_A , el número total de eslabones que irradian las personas del grupo **A** y otra que estima N_A , el número de personas en el grupo **A**.

Se escribe el estimador basado en Hansen-Hurwitz para \widehat{D}_A con el superíndice *hh*.

$$\widehat{D}_A^{hh} = \frac{\widehat{R}_A}{\widehat{N}_A} = \frac{\frac{1}{n_A} \sum_{i=1}^{n_A} \frac{1}{p_i} \cdot d_i}{\frac{1}{n_A} \sum_{i=1}^{n_A} \frac{1}{p_i}} \quad (2.76)$$

Donde p_i es la probabilidad de que la persona *i* sea seleccionada en un selección específica.

En la ecuación (2.76) las probabilidades de selección p_i son desconocidas. Sin embargo, porque las personas son seleccionadas con probabilidad proporcional al grado, se sabe que las probabilidades relativas de selección para dos nodos, *i* y *k*, serán

$$\frac{p_k}{p_i} = \frac{d_k}{d_i} \quad \forall i, k \quad (2.77)$$

Así, para cada persona se puede reescribir su probabilidad de selección, en términos de una selección de referencia de persona *k*. Por lo tanto, utilizando la ecuación (2.76) para reescribir la ecuación (2.75), se obtiene

⁴⁸ Estimación de Hansen-Hurwitz es un procedimiento estándar cuando se estima a partir de datos donde el muestreo es con reemplazo y las unidades cuentan con probabilidades desiguales de ser seleccionadas (Hansen y Hurwitz, 1943; Cochran, 1977; Brewer y Hanif, 1983). La idea básica, similar al estimador de Horvitz-Thompson, es que el procedimiento ponderado cada elemento de la muestra por la inversa de su probabilidad de selección. Es decir, las unidades con pequeña probabilidad de ser seleccionadas cuentan más.

$$\widehat{D}_A^{hh} = \frac{\frac{1}{n_A} \sum_{i=1}^{n_A} \frac{d_k}{d_i \cdot p_k} \cdot d_i}{\frac{1}{n_A} \sum_{i=1}^{n_A} \frac{d_k}{d_i \cdot p_k}} \quad (2.78)$$

Dado que $\frac{d_k}{p_k}$ es constante, se puede eliminarlo del numerador y denominador. Además, se cancela el término $\frac{1}{n_A}$, y queda

$$\widehat{D}_A^{hh} = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}} \quad (2.79)$$

Se debe tener en cuenta que la ecuación (2.79) no requiere las probabilidades de selección desconocidas, p_i , para cada nodo. Se requiere sólo la información que se recoge en la muestra.

Las ecuaciones (2.75) y (2.79) son idénticas. Es decir, se puede ver que estos dos enfoques muy diferentes - uno basado en las distribuciones del grado y otro basado en los estimadores Hansen-Hurwitz - llevan al mismo estimador. Dado que estos estimadores son iguales, se quitarán los superíndices *hh* y *dist* y se referirá al estimador como D_A .

El numerador y el denominador de \widehat{D}_A son estimadores Hansen-Hurwitz que se sabe son **insesgados**⁴⁹. También es el caso de que el cociente de dos estimadores insesgados son asintóticamente insesgados con un sesgo en el orden de n^{-1} , donde n es el tamaño de la muestra⁵⁰. Como n_A , es el tamaño de la muestra en el grupo **A**, se hace grande, $E[\widehat{D}_A] \rightarrow D_A$. Por lo general, este sesgo se considera insignificante en muestras de tamaño moderado⁵¹.

Se ha demostrado que dos enfoques diferentes para estimar la media del grado para el grupo **A** son equivalentes y que este estimador, \widehat{D}_A , es asintóticamente insesgado (ecuaciones 2.71 a la 2.79), porque es una razón de estimadores Hansen-Hurwitz y esta razón es insesgada.

2.4.3.3. Estimación de la proporción poblacional (P)

Al estimar cierta información sobre la red social que conecta a la población, se puede usar esa información para estimar la P_A y P_B .

Conectando las ecuaciones (2.78) y (2.69) en las ecuaciones que expresan las proporciones de la población en términos de información de la red, ecuaciones (2.58) y (2.59), ahora podemos estimar la **proporción poblacional** como

⁴⁹ Brewer y Hanif, 1983

⁵⁰ Cochran, 1977

⁵¹ Ibedem

$$\widehat{P}_a = \frac{\widehat{D}_b \cdot \widehat{S}_{ba}}{\widehat{D}_a \cdot \widehat{S}_{ab} + \widehat{D}_b \cdot \widehat{S}_{ba}} \quad (2.80)$$

Se obtiene un cociente de estimaciones asintóticamente insesgadas que también es asintóticamente insesgada. Con esto, se tiene una estimación asintóticamente insesgada de la proporción de la población con un rasgo específico basado solamente en datos recogidos durante el muestreo RDS.

2.4.4. Estimación de la varianza

El complicado diseño de RDS crea numerosos retos para la estimación de la varianza. Al aplicar los estimadores de varianza HH dados por Hansen y Hurwitz

$$\widehat{V}_{HH}(\widehat{T}_y) = \frac{1}{N^2 n(n-1)} \sum_s \left(\frac{y_i}{s_i} - \widehat{T}_y \right)^2 \quad (2.81)$$

Al estimar el valor de la media de y , se convierte en

$$\widehat{V}_{HH}(\widehat{y}) = \frac{1}{n(n-1)} \sum_s \left(\frac{y_i \widehat{D}}{d_i} - \widehat{y} \right)^2 \quad (2.82)$$

Pero como se ha señalado, el estimador RDS es sólo similar a los estimadores de HH , que se supone que cada unidad de muestra se extrae de forma independiente, de modo que fuera de algunos casos especiales este estimador de varianza tiene un performance bastante pobre. Cabe recordar que las unidades de muestra en el RDS se correlacionan porque están unidas por una cadena de reclutamiento, algo que no se considera en los estimadores HH estándar.

Un muestreo con probabilidades de selección diferentes es considerado en el estimador de la varianza anterior. Pero además, una muestra de RDS constituye una muestra de cadenas de Markov con un proceso de Monte Carlo de la red social, con la matriz de probabilidades de transición σ^X . En general, es necesario tener en cuenta esa correlación en la estimación de la varianza. Las unidades de la muestra están conectadas en una cadena de reclutamiento con los índices $1, \dots, n$, considerando dos unidades de la muestra I y J , la distancia de uno a otro dentro de la cadena de reclutamiento será $|i - j|$.

La correlación entre las unidades muestrales se reducirá con la distancia dentro de la cadena de reclutamiento. Para las variables categóricas, por aproximación de la probabilidad de que la unidad $i + 1$ es de tipo Y dado que i es de tipo X esto es σ_{xy} , como ya se menciono **σ es la matriz de probabilidades de transición** estimada del proceso simplificado de Markov, por ejemplo $\Pr[\text{Unit } i + 1 \in Y \mid \text{Unit } i \in X]$ es estimado como $\sigma_{XY} = R_{XY} / \bar{R}_X$.

Dadas dos unidades i y j , con i de tipo X , la probabilidad de que la unidad j es del tipo Y es calculada por el producto de la matriz de transición de probabilidades: $\Pr(j \in Y | i \in X)$ es estimado por $(\sigma^{|i-j|})_{XY}$.

La distancia entre las unidades muestrales i y j , $|1 - j|$, es el exponente al que la matriz de σ se eleva, lo que da lugar a una matriz cuadrada con las mismas dimensiones que σ , $N_d \times N_d$. En este contexto, el subíndice AB se refiere a los índices correspondientes a los conjuntos A y B , e indica el elemento de AB esimo de la matriz $\sigma^{|i-j|}$.

Se debe tener en cuenta que el uso de la matriz $\sigma^{|i-j|}$ es una simplificación, esto no siempre es el caso, el reclutamiento puede ser modelado como un proceso de Markov de primer orden con la probabilidad de transición σ y los nodos específicos de la probabilidad de transición σ^X son casi siempre desconocidos.

Se considera sólo el modelo de dos grupos, el RDS se basa en la siguiente cadena de Markov:

$$\sigma_{xy} = \begin{cases} 1(1-H)/N, & x, y \in A, \text{ o } x, y \in B \\ 2H/N, & x \in A, y \in B \text{ o } x \in B, y \in A \end{cases} \quad (2.83)$$

En forma de matriz

$$\sigma = \begin{bmatrix} 2(1-H)/N & 2H/N \\ 2H/N & 2(1-H)/N \end{bmatrix} \quad (2.84)$$

Para una muestra inicial X_0 elegida uniformemente del grupo A y una referencia en cadena de tamaño n , se tiene que.

$$\mathbb{E}(\hat{p}) = p + (p_A - p_B) \frac{1 - \beta_1^n}{4nH} \quad (2.85)$$

Donde $\beta_1 = 1 - 2H$; $p = \frac{p_A + p_B}{2}$ y H es la homofilia

Con el sesgo del estimador de \hat{p} como

$$\text{Sesgo}(\hat{p}) = (p_A - p_B) \frac{1 - (1 - 2c)^n}{4nc} \approx \frac{p_A - p_B}{4nc} = \frac{p_A - p_B}{2n(1 - \beta_1)} \quad (2.86)$$

Que depende de la homofilia H , la longitud de la cadena de n y la diferencia en las proporciones de infección entre los dos grupos. La brecha espectral $(1 - \beta_1 = 2H)$ captura el efecto de la estructura de la red.

Se debe tener en cuenta que esto también demuestra que a pesar de que las estimaciones RDS son asintóticamente insesgadas, no puede haber un sesgo sustancial cuando las semillas no son seleccionadas de la distribución estacionaria y el tamaño de la muestra es pequeño.

En las poblaciones con estructura comunitaria, es más probable que las personas refieran a personas que están en su mismo subgrupo social. En esta situación se gana menos información de cada recluta que si cada recluta fuera elegido al azar entre toda la población. El resultado de esta dependencia es una reducción efectiva en el tamaño de la muestra. Es decir, la varianza de los estimadores de RDS es más grande que la varianza que las estimaciones basadas en una muestra aleatoria simple del mismo tamaño nominal. Se considera el camino X_0, X_1, \dots definido en (2.79). Si $X_0 \sim \text{distribuidos estacionariamente}$, entonces la varianza de \hat{p} satisface

$$\text{Var}(\hat{p}) = \frac{p - p^2}{n} + \frac{(p_A - p_B)^2 \beta_1}{2n(1 - \beta_1)} - \frac{(p_A - p_B)^2 (\beta_1 - \beta_1^{n+1})}{2n^2(1 - \beta_1)^2} \quad (2.87)$$

El estimador de la varianza para una característica particular \hat{p} , donde $\hat{p} = \frac{p_A + p_B}{2}$, $\beta_1 = 1 - 2H$ y n es el tamaño de la muestra. Sea p_A y p_B , las proporciones dentro de cada grupo que cumplen con una característica específica, como ser infectados con VIH dentro de cada grupo⁵².

Con una covarianza definida a partir de (2.79). Supongamos X_0^1, X_1^1, \dots y X_0^2, X_1^2, \dots dos realizaciones independientes de la caminata con $X_0^1 = X_0^2 \sim \text{distribuidos estacionariamente}$. Es decir, las dos cadenas comienzan en el mismo vértice v , que se extrae de la distribución estacionaria. Entonces, para $i, j \geq 0$.

$$\text{Cov}(f_D(X_i^1), f_D(X_j^2)) = \left(\frac{p_A - p_B}{2}\right)^2 \beta_1^{i+j} \quad (2.88)$$

Donde $\beta_1 = 1 - 2H$ y

$$f_D(v_i) = \begin{cases} 1, & v_i \text{ cumple con la característica} \\ 0, & \text{en otro caso} \end{cases}$$

2.4.5. Medición de la precisión de la estimación

2.4.5.1. Intervalos de confianza

Antes de presentar el procedimiento de intervalo de confianza, primero se tiene que introducir un lenguaje con el que describir la población oculta. Una población oculta se compone de dos grupos mutuamente excluyentes y colectivamente exhaustivos denominados grupo **A** y grupo **B**. Los grupos podrían ser, por ejemplo, las personas con y sin VIH.

⁵² Goel, Salganik, 2009.

La proporción de la población en el grupo **A** se llamará P_a . Una estimación puntual de esta prevalencia es útil, pero es difícil de interpretar sin una cierta medida de la precisión de la estimación. Una forma común de describir esta precisión es con un intervalo de confianza que proporciona un rango dentro del cual se espera encontrar el valor real de la población con algún grado de certeza.

Se suele ignorar el hecho de que los datos son recogidos con un diseño muestral complejo y construir intervalos de confianza como si se tuviera una muestra aleatoria simple. Este enfoque de ignorar el diseño de la muestra, en general, hará que el muestreo RDS produzca intervalos de confianza muy pequeños. Estos intervalos de confianza son incorrectos y no son sólo una preocupación técnica; estos intervalos de confianza incorrectos pueden llevar a conclusiones incorrectas substanciales.

Con el fin de producir mejores intervalos de confianza, se desarrolla un método de autoreposición (*bootstrap*) diseñado específicamente para el muestreo RDS.

Aunque un enfoque analítico, sería preferible, los métodos de autoreposición se utilizan para estimar la varianza de los diseños muestrales complejos porque las soluciones analíticas a menudo no son posibles.

Los intervalos de confianza para RDS son estimados usando un algoritmo especializado de autoreposición. El algoritmo genera un remuestreo de las observaciones dependientes basados en la matriz de transición de la muestra. Es decir, si el 70% de los reclutamientos de tipo **A** son otros **A**'s y la actual observación es de tipo **A**, el algoritmo genera una **A** como la siguiente observación en el remuestreo con probabilidad 0,7. Este proceso continúa hasta que el remuestreo alcanza el tamaño de la muestra original. Los estimadores RDS son calculados y el proceso se repite hasta que el número especificado de remuestras sea alcanzado. Las colas del intervalo de confianza se toman de la distribución de estas estimaciones de autoreposición, el límite superior de un intervalo de confianza del 95%. En consecuencia, el algoritmo de autoreposición permite intervalos de confianza no simétricos y no proporciona una estimación directa de la varianza. Todos los estimadores e intervalos de confianza RDS que se presenta aquí se calculan utilizando el software RDSAT 5.6 con el nivel alfa 0,025 (en consonancia con un intervalo de confianza del 95%), 5,000 muestras para autoreposición y la configuración por defecto para todas las demás opciones.

Supuestos:

La prueba original de que el estimador es asintóticamente insesgado RDS depende de un conjunto de cinco supuestos⁵³.

1. Los encuestados mantienen relaciones de reciprocidad con las personas que ellos saben que son miembros de la población objetivo.

⁵³ Salganik y Heckathorn (2004), Heckathorn (2007)

2. Cada encuestado puede ser alcanzado por cualquier otro encuestado a través de una serie de vínculos de la red, es decir, la red forma un solo componente.
3. El muestreo es con reemplazo.
4. Los encuestados pueden informar con precisión su tamaño de la red personal o equivalentemente, su grado.
5. El reclutamiento de compañeros es una selección aleatoria de los compañeros del reclutador.

Los tres primeros supuestos especifican las condiciones necesarias para el RDS como un método de muestreo adecuado para una población. En primer lugar, con el fin de que el reclutamiento se produzca, los encuestados deben tener acceso a otros miembros de la población y ser capaces de identificar cuáles de sus compañeros califican para el reclutamiento. Además, las estimaciones de RDS se basan en una estructura de red en la que las relaciones son recíprocas. Formalmente, si **A** recluta a **B**, entonces debe haber una probabilidad no nula de que **B** podría haber reclutado a **A**.

En consecuencia, el diseño de la investigación RDS incluye medios para impulsar a los sujetos para reclutar a sus conocidos o amigos en lugar de extraños, recompensando el éxito de los reclutadores y haciendo los derechos de reclutar limitados.

En segundo lugar, la población se supone que forma un solo componente. En otras palabras, todos en la población objetivo deben ser accesibles a partir de un simple encuestado solo siguiendo un conjunto finito de vínculos de la red. En una red aleatoria, se forma un solo componente cuando los grados individuales son elevados en comparación con el logaritmo natural del tamaño de la población. Cuando a los encuestados se les permite reclutar no sólo a aquellos con los que tienen una relación especial, sino también a algunos amigos y conocidos que saben que son miembros de la población objetivo, entonces los grados individuales son más grandes que lo que generalmente se requiere para una red para formar un solo gran componente. Además, desde las redes sociales actuales no son totalmente al azar, un requisito mínimo es que no exista ninguna barrera social o estructural que segregue por completo a un subgrupo del resto de la población. En tercer lugar, la teoría estadística para la estimación de RDS se basa en un esquema de muestreo con reemplazo. En consecuencia, la fracción de muestreo debe seguir siendo lo suficientemente pequeña como para que un muestreo con reemplazo sea apropiado.

Los dos últimos supuestos RDS son potencialmente los más problemáticos. El supuesto cuatro requiere que los participantes ofrezcan información precisa de su tamaño red personal, una tarea que es difícil incluso para expertos en redes sociales. El supuesto cinco establece que los patrones de reclutamiento reflejan la composición de la red personal dentro de la población objetivo.

Es decir, RDS asume que los encuestados reclutarán como si se tratara de una selección al azar de sus redes personales⁵⁴, sin embargo la selección aleatoria es difícil en muchos lugares (de ahí la necesidad de tomar muestras complejas y técnicas de análisis).

Por ejemplo, los estudios sugieren que la memoria reciente de experiencia de contacto pueden influir en la accesibilidad del nombre de un compañero y por lo tanto la probabilidad de un intento de reclutamiento.

Alternativamente, los encuestados pueden reclutar al primer compañero elegible que interactúa con el encuestado.

En resumen, la estimación de RDS se basa en la idea de que los miembros de muchas poblaciones se conocen entre sí, por lo que, las conexiones sociales de un pequeño grupo de miembros puede ser seguida para producir una muestra poblacional.

2.4.5.2. Procedimiento de autoreposición

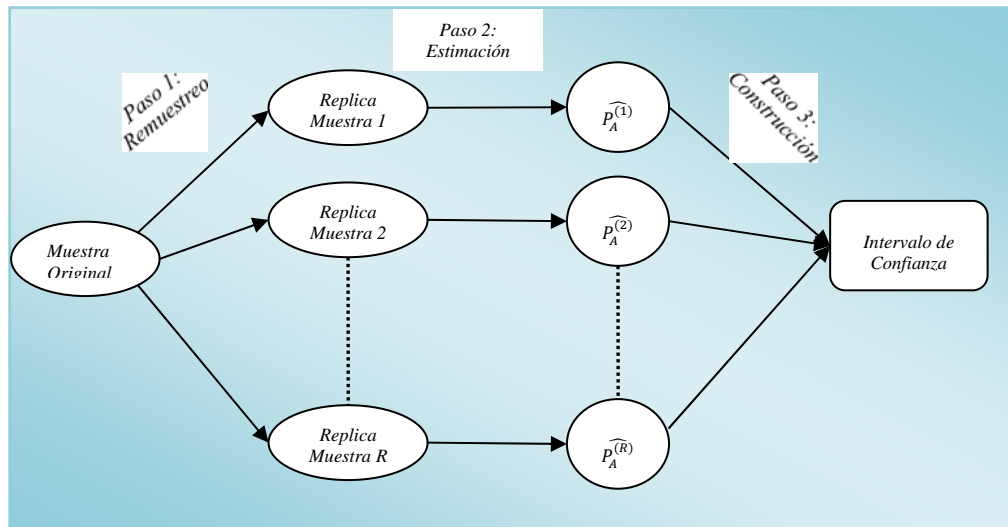
La idea general del procedimiento de autoreposición es utilizar la muestra observada para generar un conjunto de muestras idénticas. A continuación, este conjunto de réplicas de muestras se utiliza para producir un conjunto de réplicas de las estimaciones. Al examinar la variación de estas estimaciones replica, se puede construir un intervalo de confianza alrededor del punto estimado original. Este proceso de tres pasos se puede ver en la Figura 3. El primer paso en el proceso es el paso de remuestreo. En el procedimiento de autoreposición tradicional, se realiza un muestreo aleatorio con reemplazo de la muestra original hasta que la muestra repetida es del mismo tamaño que la muestra original. Este procedimiento de remuestreo está bien fundamentado teóricamente para el caso en que se recolectó la muestra original a través de un muestreo aleatorio simple. Sin embargo, como se describió anteriormente, en el muestreo RDS hay dependencias en el proceso de selección de la muestra y por ello debemos utilizar un remuestreo modificado. La modificación de la etapa de muestreo es la principal forma en que este enfoque se aparta de las técnicas de autoreposición tradicionales.

Bajo el procedimiento propuesto se divide a los miembros de la muestra en dos grupos basados en cómo se reclutó a: personas reclutadas por alguien en el grupo **A** (que llamaremos A_{rec}) y personas reclutadas por alguien en el grupo **B** (que llamaremos B_{rec}). Por ejemplo, A_{rec} podría ser el conjunto de todos los miembros de la muestra que fueron reclutados por una persona con VIH. Se debe tener en cuenta que este conjunto podría incluir tanto las personas con y sin VIH. Con el fin de imitar el proceso de muestreo real, el muestreo se inicia cuando una semilla se elige con probabilidad uniforme de toda la muestra. Luego, basándose en la pertenencia a grupos de la semilla, señalamos con el reemplazo de cualquiera de A_{rec} o B_{rec} .

⁵⁴ Heckathorn (2002)

Por ejemplo, si la semilla elegida para la muestra repetida era miembro de la muestra con VIH, que sacamos de la serie de miembros de la muestra que fueron reclutados por una persona con VIH. A continuación, examinamos la pertenencia a grupos de esta persona elegida, luego, elegimos de nuevo con reemplazo de cualquiera de A_{rec} o B_{rec} . Este proceso continúa hasta que la muestra autorepuesta es del mismo tamaño que la muestra original. En general, este remuestreo conserva algunos, pero no todas, las dependencias que existen en el muestreo RDS por el recojo de datos.

Figura 3 - Proceso de inferencia estadística



Una vez que las muestras *bootstrap* son seleccionadas, se continúa al paso 2 en la Figura 3: la fase de estimación. Aquí se utiliza el procedimiento normal de estimación de RDS en cada uno de las R muestras replicadas, para producir un conjunto de R estimaciones replicadas. Por último, en el paso 3 del procedimiento *bootstrap*, los R estimadores replicados son convertidos en un intervalo de confianza. Una forma de hacer esto sería construir un intervalo de confianza de 90% basado en la aproximación normal,

$$[\hat{P}_A - 1.65 \cdot \widehat{se}(\hat{P}_A), \hat{P}_A + 1.65 \cdot \widehat{se}(\hat{P}_A)] \quad (2.89)$$

Donde el error estándar estimado, $\widehat{se}(\hat{P}_A)$, es la desviación estándar de las estimaciones de réplicas.

Aunque este planteamiento es razonable, tiene dos desventajas principales. En primer lugar, las fuerzas de los intervalos de confianza son simétricas, lo que puede reducir la precisión y la segunda, puede producir intervalos con extremos del rango [0, 1].

2.4.5.3. Cálculo del tamaño muestral

Existe una recomendación para el uso de un tamaño de efecto de diseño de 2, que implica utilizar el doble del tamaño muestral que sería utilizado en un muestreo aleatorio simple. La información sobre los efectos del diseño se debe utilizar cuando se planifica el tamaño del estudio por muestreo RDS, o bien el tamaño de la muestra no se alcanzará.

Sin embargo, el cálculo del tamaño de la muestra en un muestreo aleatorio simple es a menudo difícil debido al carácter excesivamente general de la literatura en el análisis de poder.

Por lo tanto, se revisa los cálculos de tamaño de la muestra en dos casos específicos de mayor interés en el muestreo RDS: estimar la prevalencia de una característica con una precisión determinada y la detección de un cambio en la prevalencia en el tiempo.

$$V(\hat{P}_A) = deff \cdot \frac{P_A(1 - P_A)}{n} \quad (2.90)$$

Se puede resolver para un tamaño de muestra requerido, n , en términos del error estándar deseado, lo que da,

$$n = deff \cdot \frac{P_A(1 - P_A)}{(se(\hat{P}_A))^2} \quad (2.91)$$

Un segundo problema de interés es comparar la prevalencia de un comportamiento en dos momentos.

De manera más general, se requiere un tamaño de muestra para comparar la prevalencia en 2 poblaciones:

$$n = deff \cdot \frac{\left[\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta} \right) \cdot \sqrt{P_{A1}(1 - P_{A1}) + P_{A2}(1 - P_{A2})} \right]^2}{(P_{A2} - P_{A1})^2} \quad (2.92)$$

Dónde $Z_{1-\frac{\alpha}{2}}$ y $Z_{1-\beta}$ son los valores adecuados de la distribución normal estándar y es el efecto diseño $deff$.

Estos cálculos del tamaño de la muestra se basan en supuestos acerca de la prevalencia de las características y el efecto del diseño.

2.4.5.4. Efecto de diseño

A pesar de que el muestreo RDS produce estimadores de prevalencia insesgados y permite producir intervalos de confianza aproximados. La insesgadez de las estimaciones no garantiza que cualquier estimación particular será igual al valor real de la población. Más bien, al igual que todos los estimadores insesgados, a veces la estimación será demasiado baja o demasiado alta, sólo en promedio la estimación será igual al valor real de la población. Si las estimaciones de RDS son muy variables, entonces, incluso si son imparciales, no podrían ser de utilidad en la práctica.

El término "efecto de diseño" ha adquirido dos significados en la literatura de muestreo. El primer significado es la relación entre la varianza de la estimación correspondiente a un plan de muestreo especificados a la diferencia en el muestreo aleatorio simple (*deff*). Una definición alternativa está basada en la relación de los errores estándar (*deft*⁵⁵). Donde $\sqrt{deff} = def$

$$deff(\hat{P}_A) = \frac{V(RDS, \hat{P}_A)}{V(SRS, \hat{P}_A)} \quad (2.93)$$

Es decir, cuando $V(RDS, \hat{P}_A)$ es la varianza de RDS y $V(SRS, \hat{P}_A)$ es la varianza de un muestreo aleatorio simple. El muestreo RDS generalmente proporciona menos información, se espera que el efecto del diseño por lo general, pero no siempre, sea mayor que 1.

Los valores de los *deffs* generalmente no se conocen antes de la encuesta, como este diseño muestral no es muy común se usará los *deffs* de encuestas similares en otros países, asimismo, es valor de 2 que se asume en está es el doble de lo que un muestra aleatoria normal necesitaría.

• ⁵⁵ Deft = 1: Sin efecto de diseño muestral sobre el error estándar; Deft>1: El diseño muestral infla el error estándar de la estimación; Deft<1: El diseño muestral incrementa la eficiencia (reduce la desviación estándar) del estimador

Capítulo 3: Marco Práctico

3.1. Introducción

El método de muestreo RDS es utilizado en poblaciones muy particulares que se consideran ocultas de las cuales no se tienen marcos muestrales y por lo tanto se desconoce su verdadero tamaño.

La población que participa de este estudio tiene comportamientos estigmatizados o “mal vistos” por su preferencia sexual por lo que tiene un recelo particular para ofrecer información y por lo tanto no se tiene el tamaño real de la población y mucho menos listados o algo similar que represente un marco muestral. La preferencia sexual “no tradicional” no será aceptada por las personas que por ejemplo están casadas y con hijos, pero tienen un comportamiento bisexual, estas personas no reconocerán abiertamente una preferencia sexual diferente de la “normal” excepto a otra persona que conozca su comportamiento y que pueda identificarlo sin lugar a dudas.

Hay diferentes organizaciones de GBT que poseen listados de las personas que asisten a sus capacitaciones y que reconocen su preferencia sexual y que son el nexo con aquellas que tienen la preferencia sexual pero no la reconocen abiertamente, es a través de estos nexos que se conforma la red social y se inician los procesos de reclutamiento en cadena.

Por lo tanto y usando la teoría del marco conceptual, los estimadores que se obtengan de este estudio serán representativos de la población y para que el estudio tenga resultados válidos se debe cumplir con ciertas características:

- 1) Los informantes se reconocen los unos a los otros como miembros de la población objetivo pues, por lo tanto saben a quién seleccionar como nuevo informante;
- 2) Las redes sociales de los miembros de la población son lo suficientemente densas para garantizar cierta profundidad sociométrica, y
- 3) La población no está segmentada en subgrupos, por lo tanto las olas o encadenamientos que se generan a partir de los primeros informantes no quedan encapsuladas en los subgrupos.

Figura 4 - Casos de VIH/SIDA 1984 - 2011



El estudio es aplicado a la población de hombres que tienen sexo con hombres en la ciudad de Santa Cruz de la Sierra. Se realiza la investigación acerca de los comportamientos y actitudes relacionadas con el VIH. En Bolivia, esta es una de las poblaciones en alto riesgo de contraer VIH, por lo tanto se hace necesario tener un seguimiento epidemiológico de esta población.

De acuerdo a datos del Programa Nacional de ITS/VIH/SIDA, hasta diciembre 2011 existían 5,184 casos reportados de VIH/SIDA, de los cuales aproximadamente el 60% se presenta en Santa Cruz, por lo tanto, el contexto geográfico es el apropiado para la investigación.

Para probar los elementos más importantes en cuanto a la puesta en práctica de este modelo de muestreo se eligieron 2 variables que ayudaran a explicar el proceso de estimación dirigida a la inferencia estadística que es el eje de la presente investigación.

Estas variables son:

- Orientación sexual
- Cobro por relaciones sexuales

La primera variable ha sido elegida por un aspecto de distribución poblacional al interior del grupo de la población HSH y la segunda variable ha sido elegida por estar altamente relacionada a un comportamiento riesgoso con respecto al VIH.

3.2. Características de la muestra y selección de los entrevistados

La muestra está dividida en 2 grupos, el primero de las semillas que han sido seleccionadas de manera dirigida y el resto de la muestra que fue conseguido a través de la metodología RDS. Se mostrará las características de los 2 grupos, sin embargo para el análisis final se incluye a ambos grupos como uno solo.

3.2.1. Calculo del tamaño de la muestra

Para el cálculo del tamaño de la muestra se baso en la potencia de la muestra, teniendo en cuenta que el número final era realmente desconocido, se tomo una medida mínima que permita satisfacer la potencia de la muestra, para ello se uso (2.92).

$$n = def f \cdot \frac{\left[\left(z_{1-\frac{\alpha}{2}} + Z_{1-\beta} \right) \cdot \sqrt{P_{A1}(1 - P_{A1}) + P_{A2}(1 - P_{A2})} \right]^2}{(P_{A2} - P_{A1})^2}$$

Y se obtuvo, un tamaño de 402 entrevistas para lograr un poder estadístico del 80%, que es en definitiva la capacidad que tiene el estudio para encontrar diferencias si es que realmente las hay, es un paso fundamental tanto en la fase de diseño como en la interpretación y discusión de sus resultados.

A la hora del diseño, por tanto, se establece la magnitud mínima de la diferencia o asociación que se considere de relevancia clínica, así como el poder estadístico que se desea para el estudio y, de acuerdo con ello, calcular el tamaño de la muestra necesaria. Tras realizar el análisis estadístico, cuando se dice que no existe evidencia de que A se asocie con B o sea diferente de B, se debe cuestionar antes de nada si la ausencia de significación estadística indica realmente que no existe una diferencia o asociación clínicamente relevante, o simplemente que no se dispone de suficiente número de encuestas para obtener hallazgos significativos.

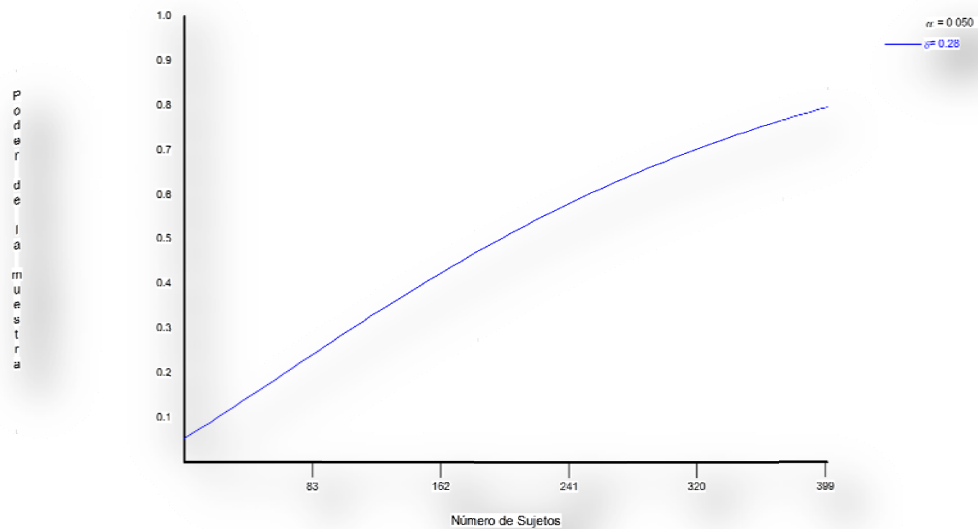
Tanto si los hallazgos son estadísticamente significativos como si no lo son, la estimación de intervalos de confianza pueden también facilitar la interpretación de los resultados en términos de magnitud y relevancia clínica, proporcionándonos una idea de la precisión con la que se ha efectuado al estimación, de la magnitud y de la dirección del efecto.

De este modo, los intervalos de confianza nos permiten tener una idea acerca del poder estadístico de un estudio y, por tanto, de la credibilidad de la ausencia de hallazgos significativos.

Si bien existe una diferencia entre el total de esperado de la muestra que es de 402 y en realidad se obtuvo 373 encuestas validas, que incluye a las semillas, se realizó el cálculo asociado al poder estadístico alcanzado con este valor, se encontró a través de:

$$z_{1-\beta} = \frac{|p_1 - p_2|\sqrt{n} - z_{1-\alpha/2}\sqrt{2p(1-p)}}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}} \quad (3.1)$$

Que el valor de poder estadístico real es de 77%, que permite afirmar que existe un suficiente tamaño de muestra para que en la comparación se pueda afirmar que los cambios obtenidos son debidos a la intervención y no ocurrieron por casualidad.



El número calculado tuvo que sufrir ajustes, dado el desconocimiento de la dimensión real de la población, por lo cual se utilizó este método para el cálculo del tamaño de la muestra.

3.2.2. Descripción de las semillas y el proceso de reclutamiento

Un total de 17 semillas fueron seleccionadas para el estudio, las mismas fueron elegidas a partir de organizaciones GBT que sugirieron a estas personas en base a procesos de trabajo, participación, conocimiento y aprendizaje que tuvieron con la organización. Adicionalmente a esto se incluye a un grupo de 3 personas seleccionadas externas a las organizaciones GBT, destinadas a mejorar el alcance al grupo de travestis y por el alto nivel de participación o motivación que presentaron estas personas, además de haber reportado un alto grado en cuanto al tamaño de sus redes sociales y su influencia en la comunidad gay.

Para cumplir con los requisitos del RDS se obtuvo información acerca del tamaño de las redes sociales asociadas a cada uno de las semillas y también se construyó un grupo de variables tendientes a conocer con la mayor precisión posible el *grado* de cada entrevistado y se utilizaron las siguientes preguntas para determinar el mismo:

- ¿Conoce a personas que se identifiquen como gay, bisexuales, travestis o trans?
- ¿Usted piensa que ellos han tenido relaciones sexuales al menos una vez con un hombre?,
- ¿Estas personas viven en Santa Cruz?,
- ¿Sabe cómo contactarse con ellos?,
- ¿Ha hablado con ellos en el último mes?,
- ¿Cuántas de estas personas ud. piensa que han tenido relaciones sexuales a cambio de dinero?

Estas preguntas eran consideradas como selectoras de los entrevistados, la última pregunta no era considerada como excluyente para el cálculo del tamaño del grado, a diferencia del resto.

Las características principales de las semillas se las muestra en la Tabla 1.

Tabla 1 - Características de las semillas participantes en el estudio (n=17)

	N
Edad	
18-25	7
26-35	6
36 y más	4
Ocupación Principal	
Trabajo regular en una empresa/ONG	11
Trabaja en su negocio propio	4
Trabajador sexual	2
Orientación Sexual	
Gay u homosexual	9
Bisexual	5
Travesti	3
Ha realizado trabajo sexual	
Sí	5
No	12

La población objetivo está formada por hombres que han tenido relaciones sexuales con otros hombres en los últimos 12 meses y que son reconocidos por sus pares y por el reclutador como hombres que tienen sexo con hombres.

La distribución de las semillas, si bien no es aleatoria, se busca que tenga representatividad en cuanto a la capacidad de reclutamiento en las diferentes sub poblaciones que existen, estas sub poblaciones están caracterizadas por la orientación sexual de los entrevistados.

Cada variable debe ser considerada respecto a sus categorías al momento de realizar las preguntas iniciales para la medida del grado.

Si los entrevistados, no dan información precisa sobre el tamaño de la red del subgrupo no se puede realizar el proceso de inferencia, porque no habrá información de las probabilidades de selección.

Se tomó como promedio de tamaño de red a 22 personas con un rango intercuartilico de 6 – 30 para la variable orientación sexual y para la variable de trabajo sexual se tomo como valor medio 9 personas con un rango intercuartilico 2 - 10.

Figura 5- Distribución del grado, orientación sexual

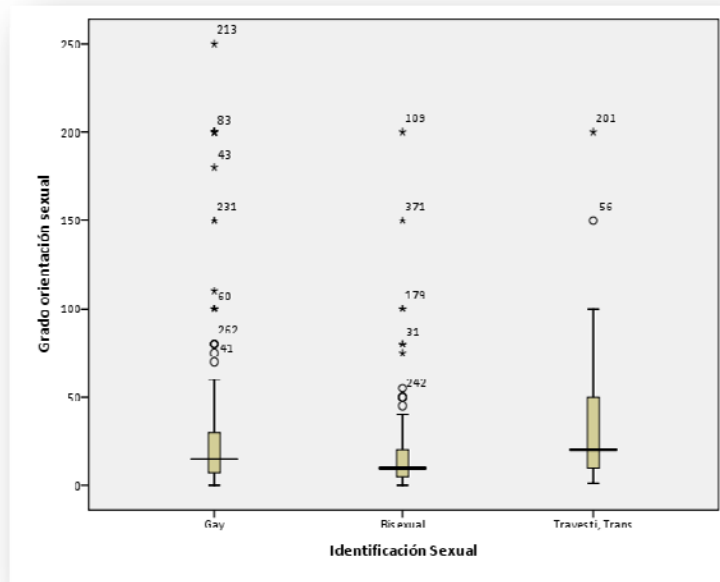
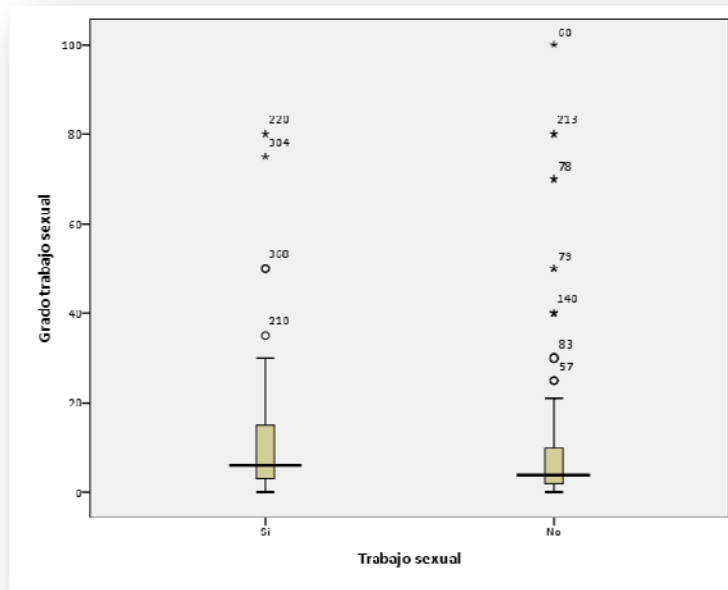


Figura 6 - Distribución del grado, trabajo sexual



3.2.3. Característica de la muestra

El procesamiento de la información fue realizado en RDSAT 5.6 para el cálculo de los estimadores y para los diagramas de las redes sociales se utilizó el software NET Draw.

Las características principales de la muestra son calculadas para un total de 356 reclutados a través de las diferentes olas, en este análisis no se considera a las semillas que representan la ola 0.

La tabla 2 muestra una descripción demográfica de la muestra, cuyas características principales son:

Grupo etario principal: *18 a 25 años*

Nivel de instrucción predominante: *secundaria a universitario incompleto*

Ingresos: *2 grupos representativos principales, aquellos que no tienen ningún tipo de ingreso y de los que tienen ingresos entre 681 a 3000 B\$.*

Orientación sexual: *predominantemente Gay*

Trabajo sexual: *22%*

Consideraciones adicionales

Un aspecto importante de resaltar es la edad de los entrevistados con el 71% en el grupo de 18 a 25 años, hecho que causo ciertos inconvenientes al momento de ir al centro de salud especializado.

Como se puede ver en la Tabla 2, durante la ola 6 se tuvo un descenso en el número de personas reclutadas, lo que hizo suponer en primera instancia que la profundidad sociométrica de las redes no era la adecuada y se había sobreestimado el número probable de entrevistados, teniendo en cuenta que el número de encuestados esperado estaba alrededor de los 400 y en la sexta ola aún faltaba aproximadamente un 30% de ese esperado. Para subsanar este impase se realizó un cambio en la estrategia de reclutamiento, implementando equipos móviles de entrevistadores, lo que genero un aumento a partir de la ola 7 que permitió acercarse al tamaño de muestra previsto.

Tabla 2 – Características socio demográficas de los entrevistados (n=356 no incluye semillas)

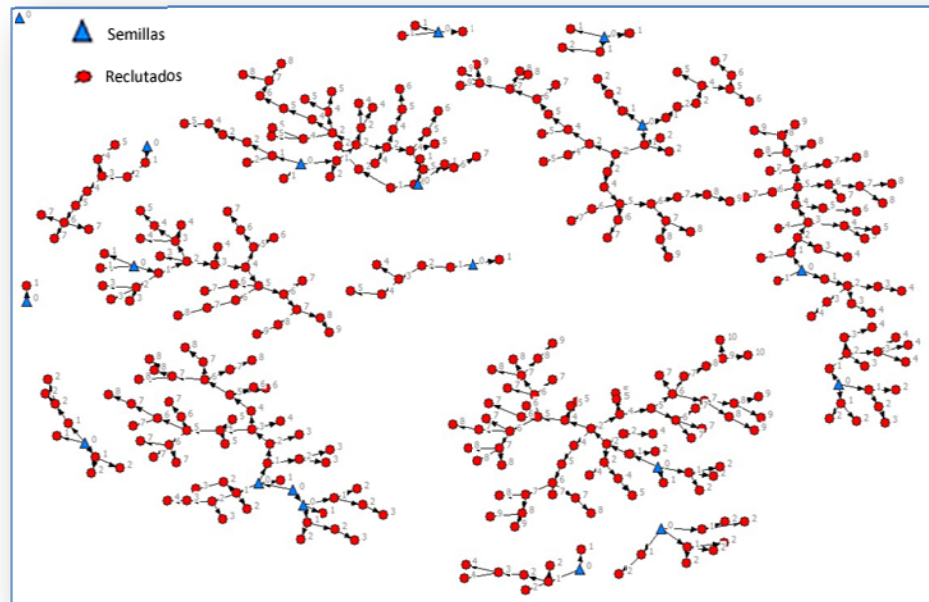
	%
Reclutamiento por ola	
Ola 1	11.5
Ola 2	12.9
Ola 3	11.5
Ola 4	14.0
Ola 5	10.7
Ola 6	9.0
Ola 7	12.6
Ola 8	11.0
Ola 9	6.2
Ola 10	0.6
Edad (en años)	
18-25	71.2
26-35	22.9
36 y más	5.9
Nivel de escolaridad	
Primaria	4.8
Secundaria	32.5
Técnico medio	4.2
Técnico superior	6.2
Universitaria incompleta	40.1
Universitaria completa	11.6
No sabe	0.3
No responde	0.3
Ingresos reportados	
No tiene ingresos	21.5
Menos de Bs. 680	6.5
Bs 681 a Bs 1400	44.9
Bs 1,401 a Bs 3,000	21.2
Mas de Bs 3,000	4.2
No responde	1.7
Orientación sexual	
Gay	49.4
Bisexual	36.7
Travesti, Trans	13.8
Trabajo sexual	
Sí	21.8
No	78.2

3.2.4. Caracterización de las redes sociales

Dentro del grupo de entrevistados, las redes sociales existentes son varias y se ha podido comprobar que la densidad sociométrica de estas fue la esperada, Figura 6.

Algunas de estas redes alcanzaron a más de 70 miembros reclutados, lo que implica una profundidad sociométrica alta, el promedio de reclutados por red es de 20 personas, que está directamente asociado al grado informado *a priori* por los entrevistados.

Figura 7 - Patrón de reclutamiento del estudio



Fuente: Elaboración Propia

La investigación produjo 10 olas, además de la ola inicial denominada ola 0, en estas 10 olas se entrevistó a 356 personas, distribuidas como se muestra en la Tabla 2 y la configuración de las redes sociales pueden observarse en la Figura 6.

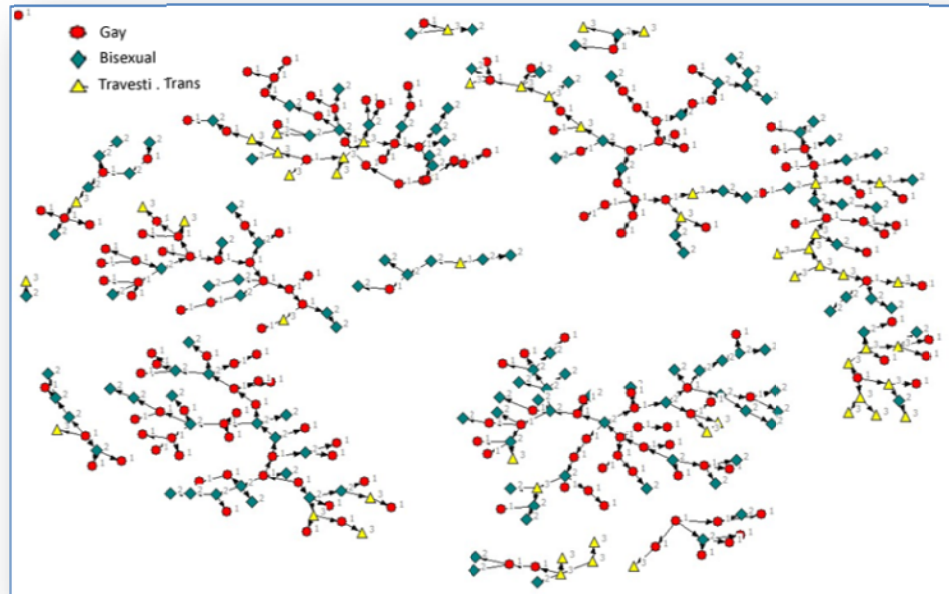
La representación gráfica muestra a las semillas y como estas produjeron el resto de los reclutados, se puede observar la participación a partir de este grupo inicial. Durante el relevamiento de información, solo una de las semillas no produjo entrevistados y para el análisis fue "aislada".

La conformación de las redes sociales de acuerdo a la orientación sexual de sus miembros se muestra en la Figura 7, de acuerdo al gráfico no se puede observar que la orientación sexual presente algún grado de endogamia o exogamia, más adelante se verificará el grado de homofilia que presenta esta variable.

Vemos en la Figura 8, que representa el proceso de reclutamiento de la variable trabajo sexual, que existen periodos u olas dentro de las redes sociales que el reclutado y el reclutador son pares, sin embargo los nodos de salida o de entrada provienen de personas no trabajadoras sexuales y no necesariamente de trabajadores sexuales y viceversa, una vez más se hace necesario el análisis de la homofilia para esta variable.

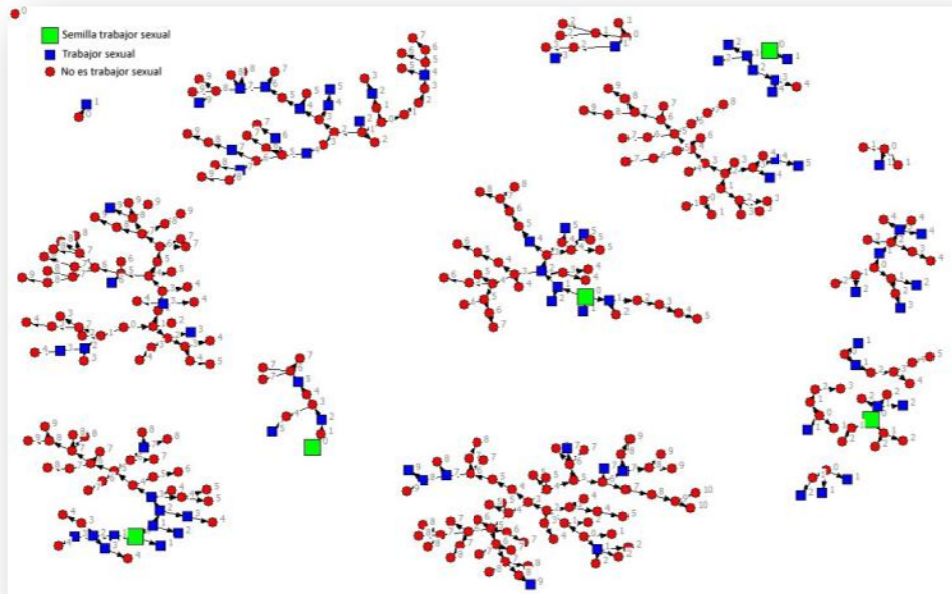
El proceso de reclutamiento es dirigido a personas que cumplen con características particulares, si estas características no se cumplen los reclutados no son contabilizados dentro del tamaño de la muestra y no son válidos para el análisis ni cualquier otro procedimiento. Esta es una ventaja con respecto a otros sistemas de muestreo porque no se tienen procesos de rechazo.

Figura 8- Estructura de las redes sociales, orientación sexual



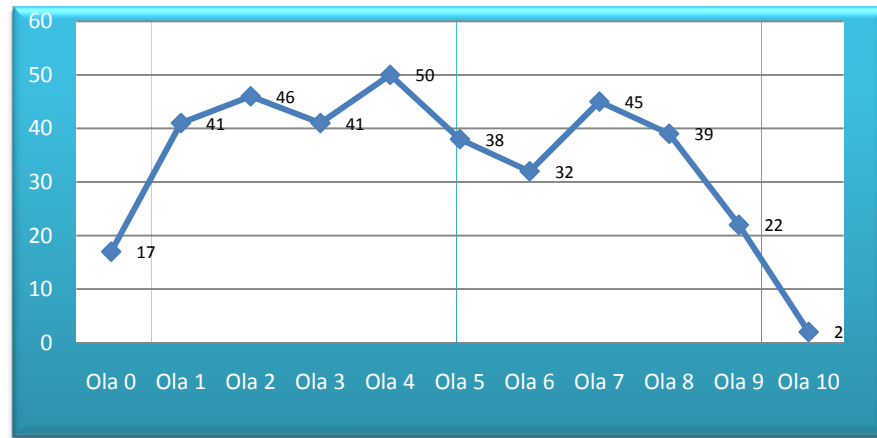
Fuente: Elaboración Propia

Figura 9 - Estructura de las redes sociales según trabajo sexual



Fuente: Elaboración Propia

Figura 10 –Número de entrevistados en cada una de las olas



Fuente: Elaboración Propia

El número de olas recomendables es variable. La muestra tiende a estabilizarse definitivamente entre la tercera y la cuarta ola. Aunque se aprecie una estabilidad de las categorías grupales con un número menor de olas se aconseja alcanzar diez o más para asegurar que la muestra obtenida se aproxima realmente a las características de la población objetivo⁵⁶.

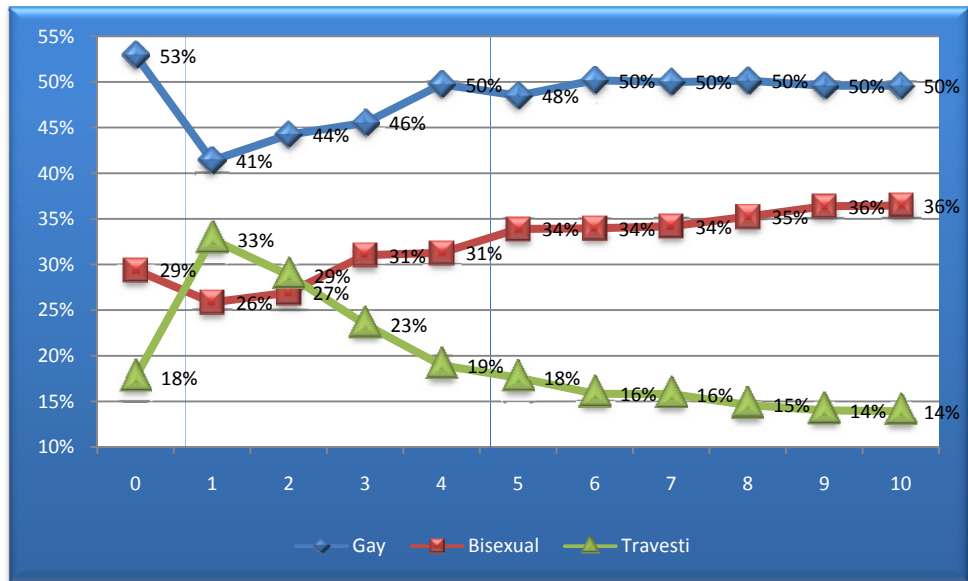
Al ser uno de los objetivos del estudio la comprensión del funcionamiento del RDS, se utilizaron sólo dos categorías muy sencillas para probar los estimadores correspondientes: la orientación sexual con las categorías de gay, bisexual y travesti y la segunda variable, trabajo sexual en los últimos 12 meses con las categorías Sí y No. En las Figuras 7 y 8 se aprecia la dinámica de reclutamiento en cada una de las categorías y la consecución de su estabilidad.

⁵⁶ Heckathorn, 2002

Tabla 3 - Distribución de los entrevistados, orientación sexual (n=373, incluye semillas)

	Gay	Bisexual	Travesti	Total
Ola 0 (semillas)	9	5	3	17
Ola 1	15	10	16	41
Ola 2	22	13	11	46
Ola 3	20	17	4	41
Ola 4	31	16	3	50
Ola 5	16	18	4	38
Ola 6	20	11	1	32
Ola 7	22	16	7	45
Ola 8	20	17	2	39
Ola 9	9	12	1	22
Ola 10	1	1	0	2
Total	185	136	52	373

Figura 11 - Evolución del % de entrevistados según orientación sexual



Fuente: Elaboración Propia

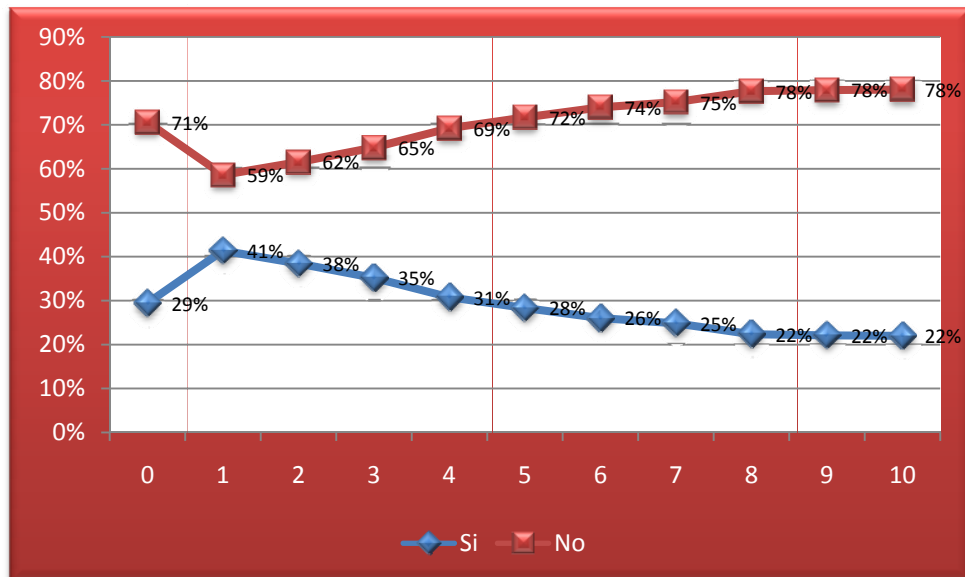
La preocupación inicial de que las semillas al ser seleccionadas no aleatoriamente pueden causar un sesgo en la distribución final de la muestra según orientación sexual se refuta a partir de la Figura 11, donde se puede observar en los diferentes grupos la variación en el reclutamiento, alcanzan su punto de equilibrio en diferentes olas, el grupo gay alcanza el equilibrio de manera más acelerada que los otros dos, el equilibrio se da en la ola 4, en el grupo de bisexuales se alcanza el equilibrio en la ola 5, sin embargo en el tercer grupo de travestis el equilibrio recién se alcanza en la ola 6 a partir de la cual se mantiene relativamente estable con una variación mínima que se mantiene por las 2 últimas olas.

Tabla 4 - Distribución de los entrevistados según trabajo sexual (incluye semillas) (n=373, incluye semillas)

	Si	No	Total
Ola 0	5	12	17
Ola 1	19	22	41
Ola 2	16	30	46
Ola 3	11	30	41
Ola 4	9	41	50
Ola 5	6	32	38
Ola 6	3	29	32
Ola 7	8	37	45
Ola 8	1	38	39
Ola 9	4	18	22
Ola 10	0	2	2
Total	82	291	373

En la variable de trabajo sexual, donde 5 semillas iniciales, trabajadores sexuales, generan una ola 1 con el 71% de reclutados trabajadores sexuales, esta variable a partir de la ola 7 halla su equilibrio, por lo tanto la selección de nuevos reclutados es independiente de las semillas, el proceso de reclutamiento se puede observar en la Figura 12.

Figura 12 - Evolución del % de entrevistados según comportamiento de trabajo sexual (n=373, incluye semillas)



Fuente: Elaboración Propia

3.3. Estimación RDS

Una muestra de referencia en cadena, en sí misma, por lo general no es representativa de la población objetivo. RDS es una adaptación del método de referencia en cadena destinado a “corregir” la representatividad.

Una muestra de RDS de una población oculta se basa en varios supuestos y requisitos mencionados en el marco teórico.

Con el fin de ayudar de aplicar de mejor manera la metodología RDS, existe un software destinado a analizar datos RDS. El Respondent Driven Sampling Analysis Tool (RDSAT)⁵⁷ que incluye las siguientes características:

Estimación de parámetros poblacionales importantes, tales como:

- Proporciones de población de los grupos definidos por el usuario
- Promedio tamaños red personal
- Homofilia
- Medidas de significación estadística de las estimaciones de población

Los datos de visualización y edición

- Exportación de resultados de análisis a formato HTML
- Visualización de datos y gráficos
- Análisis de los grupos definidos por un punto de ruptura de una variable continua

Las estimaciones realizadas de los datos RDS utilizando RDSAT pueden ser referidas como las **estimaciones de proporciones poblacionales (P)**. Las P son un ajuste de la diferencia entre la composición de la muestra y la composición de la población objetivo (las proporciones muestrales deben representar a la población). Las P se utilizan para hacer inferencias sobre la población objetivo.

3.3.1. Detalles de la estimación de RDS

Para producir las P, RDSAT utiliza datos acerca de los patrones de reclutamiento de todos los participantes en la muestra para estimar la probabilidad de que una persona con una orientación sexual gay reclutará a una persona bisexual en lugar de otra persona gay. Esto se llama probabilidad de transición. RDSAT estima la probabilidad que una persona de un tipo puede reclutar a una persona del mismo tipo o de otro tipo y la combina con la media del tamaño de la red, también conocida como **grado**, obtenida a partir de los encuestados cuando estos proporcionan la información de tamaño de la red. Estas medidas se combinan para estimar la proporción de la población perteneciente a una, dos (o más) subtipos distintos (por ejemplo, gay, bisexuales y travestis).

El análisis para la variable de orientación sexual, con las personas gay (grupo a), bisexuales (grupo b) y travestis (grupo c).

La variable de trabajo sexual, que tiene 2 grupos, el primero aquellos hombres que realizaron trabajo sexual en los últimos 12 meses (grupo a) y aquellos que no son trabajadores sexuales (grupo b).

⁵⁷ Volz, E., Wejnert, C., Degani, I., and Heckathorn, D. D. 2007 Respondent-Driven Sampling Analysis Tool (RDSAT) Version 5.6. Ithaca, NY: Cornell University.

Se desarrolla el mismo tipo de estimaciones para ambas variables, esto permite tener estimaciones poblacionales en cuanto a composición poblacional (variable orientación sexual) y comportamiento (variable trabajo sexual).

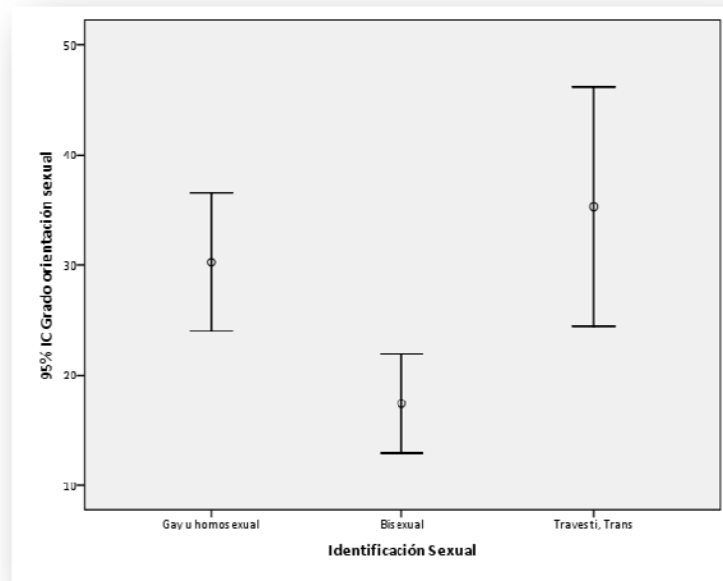
Tabla 5 -Estimadores RDS, variable en estudio orientación sexual

	Gay	Bisexual	Travesti	Total
Gay				
Recuento de reclutados	(1) 82	69	20	171
Proporción muestral, <i>Saa, Sab, Sac</i>	(2) 0.48	0.404	0.117	1
Bisexual				
Recuento de reclutados	(3) 70	43	13	126
Proporción muestral, <i>Sba, Sbb, Sbc</i>	(4) 0.556	0.341	0.103	2
Travesti, trans				
Recuento de reclutados	(5) 24	19	16	59
Proporción muestral, <i>Sca, Scb, Scc</i>	(6) 0.407	0.322	0.271	1
Distribución total de los reclutados	(7) 176	131	49	356
Distribución muestral, SD	(8) 0.496	0.365	0.138	1
Muestra en equilibrio, E	(9) 0.498	0.37	0.132	1
Diferencia media entre SD y E	(10)	0.4%		
Tamaño medio de la red ajustado, <i>Da, Db, Dc</i>	(11) 9.789	5.787	11.383	
Tamaño medio de la red No ajustado, <i>Da, Db, Dc</i>	(12) 30.934	17.97	35.288	
Estimación de la proporción poblacional, <i>Pa, Pb, Pc</i>	(13) 0.403	0.506	0.092	1
Homofilia, H	(14) 0.129	-0.325	0.197	
Error estándar	(15) 0.033	0.035	0.022	

Tabla 6 - Intervalos de confianza para las proporciones poblacionales *Pa, Pb, Pc* (alfa=0.05) – Orientación sexual

	Proporción poblacional <i>Pa, Pb, Pc</i>	Intervalo inferior	Intervalo superior
Gay	0.403	0.342	0.472
Bisexuales	0.506	0.433	0.567
Travestis, trans	0.092	0.055	0.140

Figura 13 - Intervalo de confianza (95%) para el grado no ajustado, orientación sexual



Fuente: Elaboración Propia

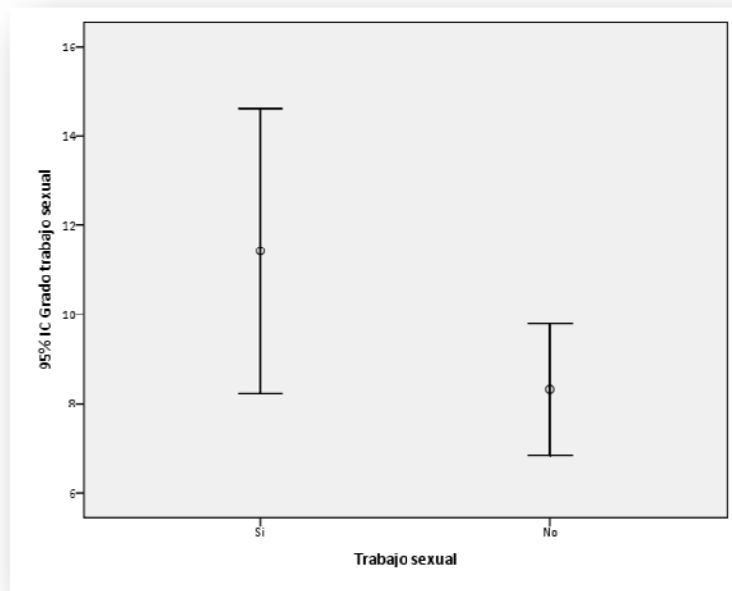
Tabla 7 - Estimadores RDS, variable en estudio trabajo sexual

		Trabajador sexual	No es trabajador sexual	Total
Trabajador sexual				
Recuento de reclutados	(1)	31	53	84
Proporción muestral, S_{aa} , S_{ab}	(2)	0.369	0.631	1
No es trabajador sexual				
Recuento de reclutados	(3)	46	226	272
Proporción muestral, S_{ba} , S_{bb}	(4)	0.169	0.831	1
Distribución total de los reclutados	(5)	77	279	356
Distribución muestral, SD	(6)	0.22	0.78	1
Muestra en equilibrio, E_a , E_b	(7)	0.211	0.789	1
Diferencia media entre SD y E	(8)	0.9%		
Tamaño medio de la red ajustado, D_a , D_b	(9)	3.917	2.934	
Tamaño medio de la red No ajustado, D_a , D_b	(10)	7.037	5.687	
Estimación de la proporción poblacional, P_a , P_b	(11)	0.167	0.833	
Homofilia, H	(12)	0.242	-0.002	
Error estándar	(13)	0.023	0.023	

Tabla 8 - Intervalos de confianza para las proporciones poblacionales P_a y P_b (alfa=0.05) – trabajo sexual

	Proporción poblacional P_a , P_b	Intervalo inferior	Intervalo superior
Trabajadores sexuales	0.167	0.122	0.214
No trabajadores sexuales	0.833	0.786	0.878

Figura 14 - Intervalo de confianza (95%) para el grado no ajustado, trabajo sexual



Fuente: Elaboración Propia

3.3.1.1. Matriz de reclutamiento

La matriz de reclutamiento describe el patrón de los reclutamientos dentro de la muestra. Para las variables en estudio se mide el número de reclutamientos de cada tipo a cada tipo. Estos datos se resumen en una matriz con el tipo del reclutador en las filas y el tipo de recluta en las columnas, en la Tabla 5 – Orientación sexual y tabla 7 – Trabajo sexual, filas 1 a 6 en la primer variable y 1 al 4 en la segunda.

Para la variable de orientación sexual, la primera columna indica que 82 de los entrevistados gay han sido seleccionados por personas pertenecientes a su mismo grupo. La segunda columna indica que 69 de los entrevistados bisexuales han sido reclutados por el grupo de personas gay y por último la penúltima columna indica que 20 de los reclutados travestis fueron reclutados por el grupo gay.

El proceso de análisis similar al anterior continúa con las líneas 3, 4 5 y 6 para la primer variable y hasta la línea 4 en el caso de la segunda variable.

El total es de 356 reclutados por las 17 semillas, el total de la muestra es de 373 personas.

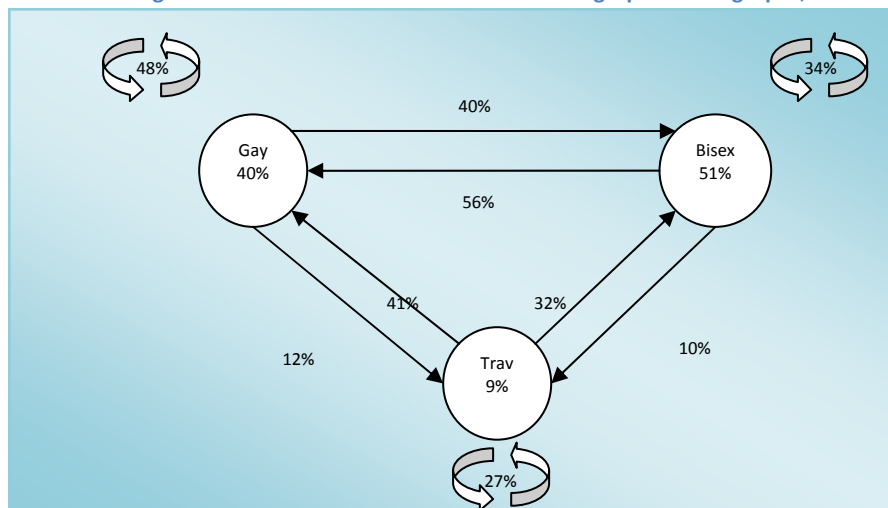
3.3.1.2. Proporción Muestral

La **proporción muestral** (S) en RDS es la razón de dos estimadores de Horvitz-Thompson, estimador estadístico insesgado y apropiado ya que asume que todos los participantes tienen diferentes probabilidades de selección⁵⁸.

También denominada probabilidades de transición, el procedimiento de pesar cada elemento de la muestra por la inversa de su probabilidad de selección para las unidades con una mínima probabilidad de ser seleccionados tiene más peso.

En otras palabras, al grupo con el tamaño de red más grande que el promedio se le asigna menos peso y al grupo con el tamaño de red media menor se le asigna más peso.

Figura 15 - Probabilidades de transición entre grupos e intragrupos, orientación sexual

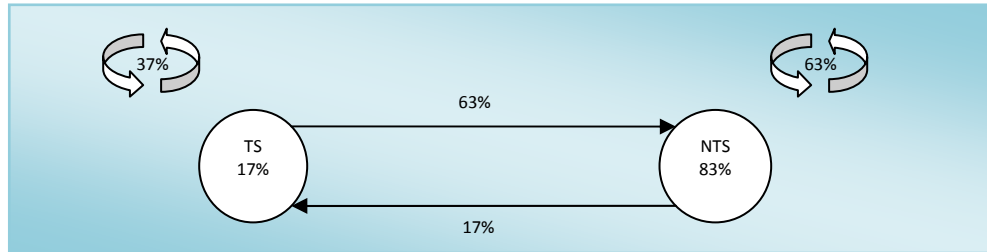


Fuente: Elaboración Propia

⁵⁸ Volz, 2008; Heckathorn, 2007; Salganik, 2006; Salganik y Heckathorn, 2003, 2004

Verificamos la existencia de una distribución estacionaria, dado que la cadena es irreducible (es decir existe una única clase de estados, y los estados que la componen son recurrentes positivos aperiódicos. Como vemos en la figura 14, los diferentes estados (Gay, bisexual, travesti) se comunican, se sabe que 2 estados que se comunican pertenecen a una misma clase de estados, como todos los estados se comunican entre sí, la cadena se denomina irreducible. Los tres estados son recurrentes y dado que la cadena tiene una cantidad finita de estados se puede afirmar que estos son recurrentes positivos.

Figura 16 - Probabilidades de transición entre grupos e intragrupos, trabajo sexual



Fuente: Elaboración Propia

3.3.1.3. Proporción poblacional

Para encontrar la P , en el caso de la variable trabajo sexual se tiene que S_{ab} es el porcentaje de hombres que son trabajadores sexuales (grupo **A**) seleccionados en el reclutamiento por hombres que NO son trabajadores sexuales (grupo **B**) y S_{ba} es el porcentaje de hombres que NO son trabajadores sexuales (grupo **A**) seleccionados en el reclutamiento por los hombres que son trabajadores sexuales. D_A y D_B son los tamaños de las redes del grupo **A** y el grupo **B** respectivamente. Utilizando las ecuaciones (2.58) y (2.59) se tiene:

$$P_a = \frac{D_b \cdot S_{ba}}{D_a \cdot S_{ab} + D_b \cdot S_{ba}} \quad (2.58)$$

$$P_b = \frac{D_a \cdot S_{ab}}{D_a \cdot S_{ab} + D_b \cdot S_{ba}} \quad (2.59)$$

Donde

P_a = proporción estimada del grupo **A**

P_b = proporción estimada del grupo **B**

S_{ab} = proporción del grupo A seleccionado por el grupo **B**

S_{ba} = proporción del grupo B seleccionado por el grupo **A**

D_a = tamaño de la red para el grupo **A**

D_b = tamaño de la red para el grupo **B**

Esta ecuación es la estimación de la proporción del grupo basado en el supuesto de que los reclutas y los reclutadores se conocen entre sí, de manera que si la persona **A** está dispuesta a reclutar a la persona **B**, entonces la persona **B** está dispuesta a reclutar a la persona **A**⁵⁹.

En las tablas 5 y 7 podemos observar los valores obtenidos para las variables en estudio, en la línea 12 para la variable orientación sexual y en la línea 10 para la variable trabajo sexual, estos valores se asumen como poblacionales presentes en la ciudad de Santa Cruz, para el grupo de HSH.

Figura 17 – Proporción poblacional, orientación sexual

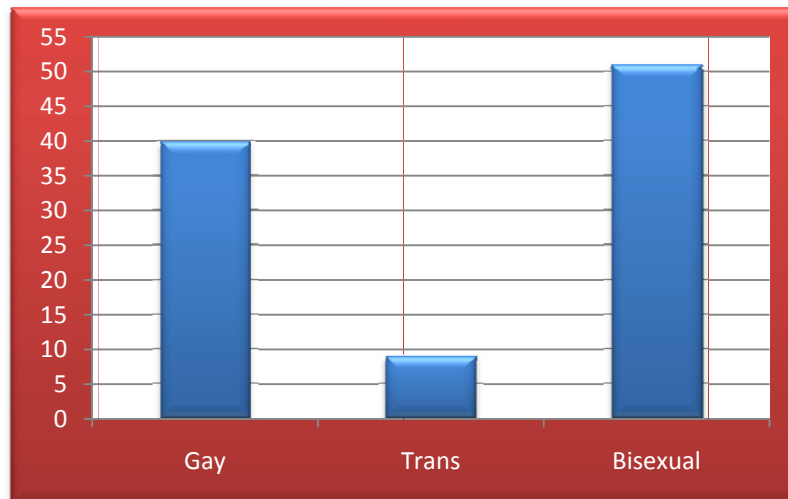
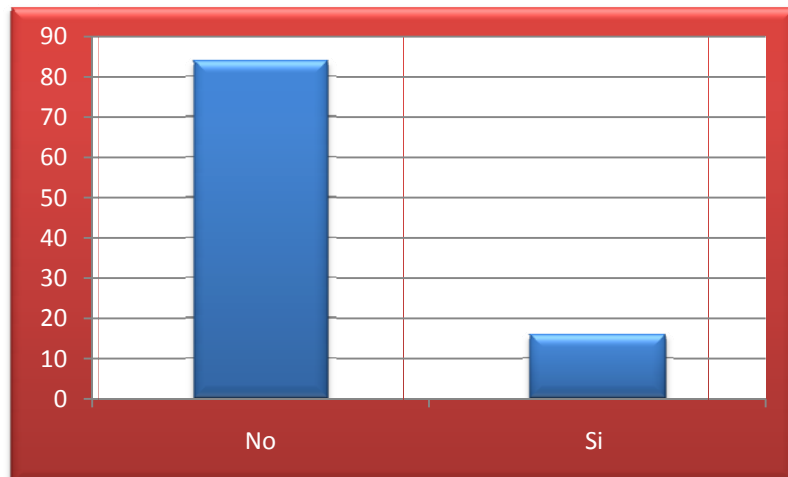
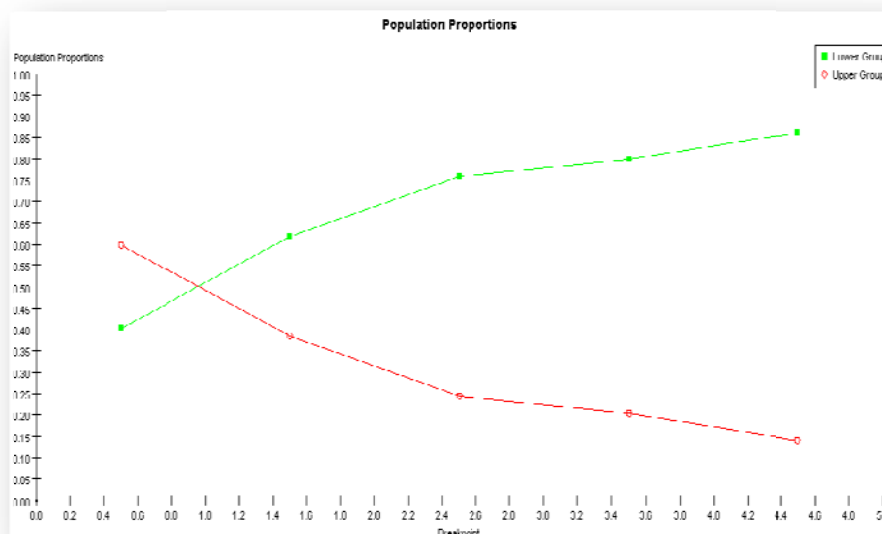


Figura 18 – Proporción poblacional, trabajo sexual



⁵⁹ Heckathorn y Rosenstein, 2002c

Figura 19 – Proporción poblacional, edad



Fuente: Elaboración Propia

Como RDS asume que el reclutamiento sigue un proceso de Markov, los cupones se distribuyen aleatoriamente en un sistema cerrado hasta alcanzar un estado de equilibrio⁶⁰. Cada ola de la cadena de reclutamiento representa un estado particular. Es decir, todas las características de los individuos en una ola de reclutamiento son estáticas en el punto en el que se alcanza la ola. Por ejemplo, en una ola de cuatro semillas, puede haber diez reclutas, de los cuales cinco son gay y cinco son bisexuales. Este estado no puede cambiar (gay y bisexual se asume que no pueden cambiar, indistinto de la orientación sexual que adopte el entrevistado posteriormente a la entrevista).

Hay dos características importantes en el proceso de Markov utilizado por RDS:

- En primer lugar, hay un número limitado de estados específicos (por ejemplo, gay/bisexual/travesti, el estado VIH-positivo o negativo, etc) que los participantes pueden asumir.
- En segundo lugar, cualquier recluta de los participantes están en función de su tipo, tal como su orientación sexual.

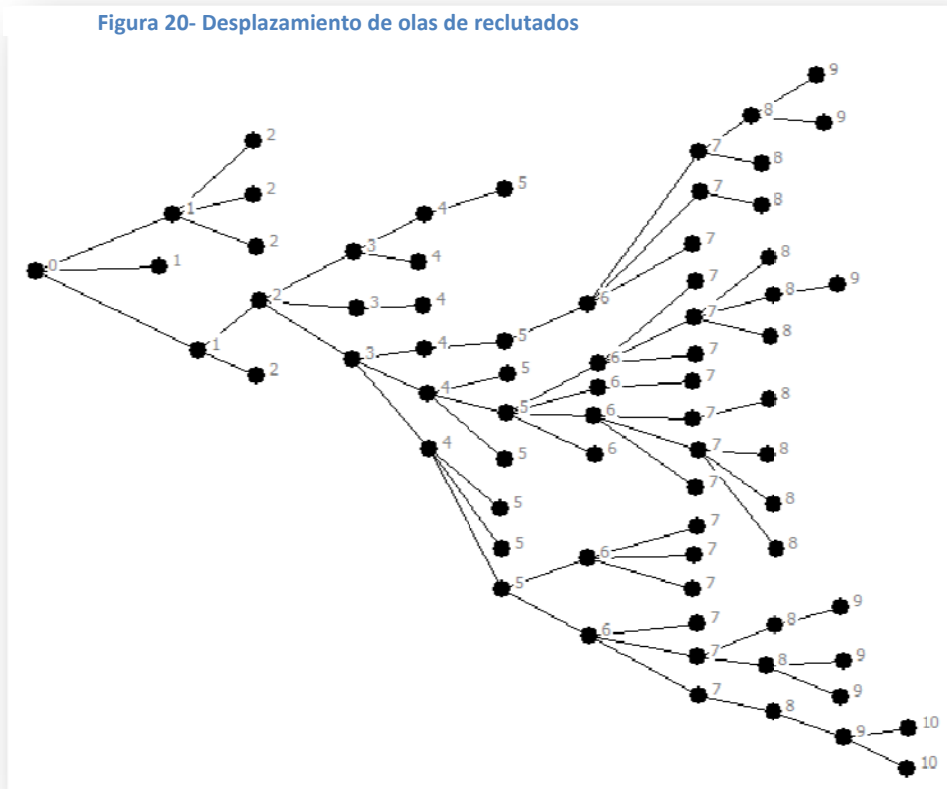
A partir de la semilla, hasta tres reclutamientos son permitidos por cada participante del estudio. Cada participante que distribuye algunos o la totalidad de sus cupones, esto genera una expansión geométrica de las olas como se muestra en la Figura 16. Cada una de estas olas representa un paso en un proceso de Markov. En cada paso (ola) las proporciones de la muestra se desplazarán hacia el crecimiento de la muestra.

⁶⁰ Goel y Salganik, 2007

El cambio de estado por el que gravitan hacia el equilibrio las proporciones de la muestra ya no se basan en que el cambio de estado de la variable (estado VIH, estado orientación sexual, etc), esto se puede ver en la Figura 5, donde las semillas tenían una determinada orientación sexual, la ola 1 en el grupo más notorio (travestis – trans), alcanzó un sobre muestreo (en esa ola) y se estabiliza en la ola 3, al convertirse el reclutamiento independiente de la semilla.

- Los participantes tienen más probabilidades de contratar a personas similares a ellos.
- En el proceso de reclutamiento RDS, los reclutas de las semillas constituyen una de las olas (ola 1).
- Los reclutados por la ola uno componen la ola dos. Los participantes de la ola dos también reclutan gente que son similares a ellos.
- Los reclutas de la ola dos forman la ola tres. Sin embargo, las características de los participantes de la ola tres, aunque similares a los de la segunda ola, comenzará a diferir en las características de los de ola uno.
- Las características de la ola cuatro, aunque similares a los de la ola tres, serán diferentes de los de las olas uno y dos, para el caso de estudio, se alcanza el equilibrio entre las olas 3 y 4, por lo tanto existe independencia del tipo de reclutados en estas olas respecto a la ola 1 y las características de las semillas.

Figura 20- Desplazamiento de olas de reclutados



Fuente: Elaboración Propia

3.3.1.4. Equilibrio de la muestra

La muestra convergerá y llegará a una estimación de la proporción de equilibrio (E) en ambas variables (orientación sexual y trabajo sexual) se alcanza entre las olas 3 y 4.

El equilibrio es el estado por el cual la variación muestral se presenta sin importar cuántas olas sean reclutadas⁶¹. Por lo tanto, las cadenas de reclutamiento deben ser suficientemente largas para alcanzar el equilibrio, para garantizar que el sesgo introducido desde la selección inicial de las semillas sea eliminado.

El número de olas requerido por las cadenas de Markov para llegar a un equilibrio de las convergencias puede ser estimado analíticamente⁶², para distintos grupos, como los trabajadores sexuales (grupo A) y los no trabajadores sexuales (grupo B).

Para calcular el equilibrio de las ecuaciones (2) y (3):

$$E_a = \frac{S_{ba}}{1 - S_{aa} + S_{ba}} \quad (2)$$

$$E_b = 1 - E_a \quad (3)$$

Donde

E_a = proporción del equilibrio del grupo A

E_b = proporción del equilibrio del grupo B

S_{ba} = proporción del grupo A seleccionado por el grupo B

S_{aa} = proporción del grupo A seleccionado por el grupo A

Esta comparación provee una indicación del sesgo en la selección inicial de las semillas. Por ejemplo, en la Tabla 7, fila 7, y representado gráficamente en la Figura 9 se puede ver que hay un aumento en el porcentaje de trabajadores sexuales en la muestra a medida que avanza a través de las olas. Entre las olas tres y cuatro se alcanza el equilibrio que es 0.21 para trabajadores sexuales y 1- E_a (0.79) para no trabajadores sexuales.

En el caso de orientación sexual el equilibrio se alcanza para el grupo A al alcanzar el valor 0.5 en el porcentaje de de entrevistados por ola con respecto a los otros dos grupos, también se alcanza este valor entre las 3 y 4, encontrando su estabilidad en este valor.

Sin embargo, dado que más personas se reclutaron más allá de las cuatro olas, la muestra no termina hasta la ola diez. Sin embargo, la S y las estimaciones E son similares indicando que la muestra cumple con los principios teóricos de un proceso de cadena de Markov (la muestra final es independiente de las características de las semillas no seleccionadas al azar).

⁶¹ Heckathorn, 1997; Salganik y Heckathorn, 2004

⁶² Heckathorn, 2002a

El factor más importante es que la muestra alcanzó un punto de equilibrio, ya que puede darse el caso que la población no tenga profundidad sociométrica y por lo tanto no se alcanza un equilibrio muestral, lo que generaría que los valores encontrados no sean útiles para inferir a la población en estudio.

Se considera una aproximación al equilibrio cuando la discrepancia es menor al 2% entre el equilibrio y la composición de la ola específica de la muestra para cada uno de los n grupos que constituyen la variable en estudio.

3.3.1.5. Reclutamiento diferencial

La matriz de reclutamiento es útil para determinar los patrones de reclutamiento diferenciales que se producen cuando algunos subgrupos sistemáticamente reclutan más personas de un subgrupo que de otro subgrupo. Si en la matriz de reclutamiento de trabajo sexual la fila con trabajadores sexuales es mayor que la fila de no trabajadores sexuales, entonces los trabajadores sexuales están reclutando más rápido que los no trabajadores sexuales.

Tabla 9 - Matriz de reclutamiento diferencial

	Trabajador sexual	No es trabajador sexual
Trabajador sexual	27.77	47.48
No es trabajador sexual	47.48	233.27

Por ejemplo, en la tabla 9 la fila para el grupo 2 (47) es mayor que la fila correspondiente al grupo 1 (27), lo que podría ser un indicio de que el grupo 2 está reclutando más rápido que el grupo 1.

El reclutamiento diferencial no afectará el equilibrio del modelo de Markov, las probabilidades de transición o el tamaño medio estimado de la red. Porque el análisis RDS depende de la distribución proporcional de los reclutas en lugar de el número de personas reclutadas. Debido a que las probabilidades de transición se basan en la distribución proporcional de los reclutas de cada grupo las probabilidades siguen siendo las mismas si todos los grupos reclutan por igual o no⁶³.

Para compensar las diferencias en el reclutamiento, RDSAT ajusta los conteos del reclutamiento, de modo que el número de reclutas de cada grupo (suma de filas) es igual al número de reclutamientos por cada grupo (suma de columnas), sin ningún cambio en el patrón de reclutamiento o tamaño de la muestra. El conteo ajustado del reclutamiento, es la probabilidad de transición multiplicada por el equilibrio de los reclutas de esa categoría y el número total de reclutamientos para todas las categorías.

⁶³ Heckathorn, 2002a

La Tabla 9 proporciona un ejemplo de los conteos de reclutamiento demográficamente ajustados o datos suavizados (ajustados por la tasa de reclutamiento por parte de las variables demográficas). El equilibrio y la proporción poblacional son calculadas como si todos los grupos reclutaran por igual.

3.3.1.6. Suavización de los datos

El análisis RDS requiere que ninguna fila o columna de la matriz de reclutamiento esté vacía. Los datos son suavizados y mejorados, llena las celdas vacías (en caso de haberlas) de la matriz de reclutamiento con la media de la diagonal.

La suavización de datos no tiene ningún efecto en tablas de dos categoría porque el problema con sobre determinación en una celda de la matriz de reclutamiento frente a otra celda vacía⁶⁴. Una de las principales ventajas de la suavización de datos es que los intervalos de confianza serán más pequeños.

Para la variable orientación sexual, la matriz de datos suavizados es:

Tabla 10 - Comparación de datos, orientación sexual

	Datos originales			Datos Suavizados		
	Gay	Bisexual	Travesti	Gay	Bisexual	Travesti
Gay	82	69	20	85.02	72.33	19.95
Bisexual	70	43	13	72.33	44.92	14.37
Travesti	24	19	16	19.95	14.37	12.77

Tabla 11 - Comparación de datos, trabajo sexual

	Datos originales		Datos Suavizados	
	Trabajador sexual	No es trabajador sexual	Trabajador sexual	No es trabajador sexual
Trabajador sexual	31	53	27.77	47.48
No es trabajador sexual	46	226	47.48	233.27

3.3.1.7. Homofilia

Homofilia (o índice de agrupación) es un estadístico que describe los patrones de mezcla en las redes. RDS es uno de los pocos métodos de muestreo que pueden proporcionar mediciones de la homofilia. Describe la probabilidad de una persona que realiza trabajo sexual refiera exitosamente a otra persona trabajador sexual de una población de individuos trabajadores y no trabajadores sexuales.

La homofilia puede ser negativa o positiva, en un rango de -1 hasta 1 dependiendo de si un individuo atribuye preferentemente a los demás de su propio tipo, o, alternativamente, los evita. Cuando la homofilia es cero para todos los grupos, el

⁶⁴ Heckathorn, 2002a

equilibrio será idéntico a la real proporción poblacional⁶⁵. Si los individuos trabajadores sexuales conforman el 16% de la población que está siendo muestreada, la teoría de Markov dice que si cualquier persona trabajador sexual se escoge en esa población, 16%, de sus conexiones también son trabajadores sexuales.

Sin embargo, esto no suele ser el caso, ya que con frecuencia las personas tienen más conexiones con gente como ellos que a las personas que son diferentes⁶⁶.

Si la homofilia es cero, entonces la proporción muestral podría igualar a la proporción poblacional. Homofilia cero, indica que los individuos de un determinado tipo, trabajadores sexuales, no tienen preferencia para reclutar ni a trabajadores sexuales ni a no trabajadores sexuales. Existe una tendencia al reclutamiento en el interior del grupo, existe una predisposición de los trabajadores sexuales a reclutar a otros trabajadores sexuales.

Caso contrario ocurre en la variable orientación sexual, donde la homofilia es negativa y relativamente alta, es decir existe heterofilia, donde los bisexuales no reclutan a otros bisexuales, sino más bien están altamente relacionados a los otros grupos.

Detalles del cálculo de la Homofilia. De las ecuaciones (7), (8) y (9) se tiene

$$H_a = \frac{D_a - (S_{ba}D_b) - (S_{ab}D_a)}{D_a} \left. \vphantom{H_a} \right\} \text{Si } S_{aa} \geq P_a \quad (7)$$

$$H_a = \frac{D_a - (S_{ba}D_b) - (S_{ab}D_a)}{S_{ba}D_aS_{ab}} \left. \vphantom{H_a} \right\} \text{Si } S_{aa} < P_a \quad (8)$$

$$H_b = \frac{S_{ba} - P_a}{-P_a} \quad (9)$$

Donde

H_a = homofilia del grupo A

H_b = homofilia del grupo B

N_a = es la media del tamaño de la red para el grupo A

N_b = es la media del tamaño de la red para el grupo B

S_{ab} = proporción del grupo B seleccionado por el grupo A

S_{ba} = proporción del grupo A seleccionado por el grupo B

S_{aa} = proporción del grupo A seleccionado por el grupo A

P_a = Tamaño proporcional del grupo A

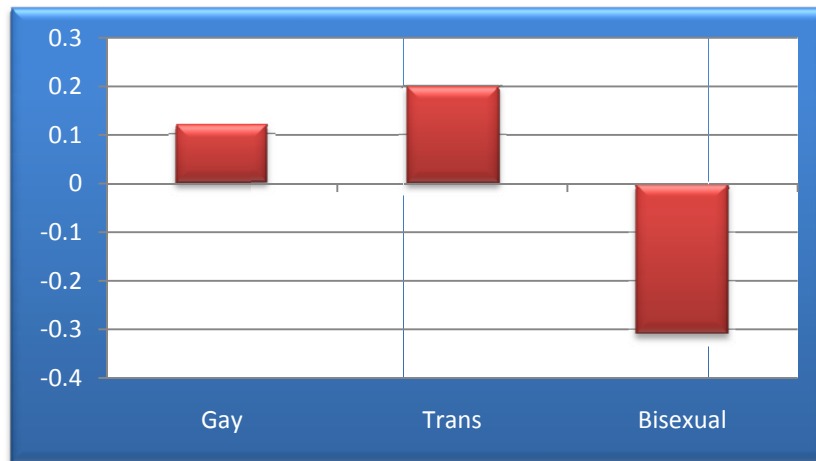
⁶⁵ Heckathorn, 1997; 2002a; McPherson et al., 2001

⁶⁶ Heckathorn, 1997; 2002a

Para la variable orientación sexual, de la Tabla 5, línea 14 se tiene que la homofilia para el grupo gay es 0.129, valor no muy alto, que muestra un proceso de reclutamiento similar intra grupo y afuera del grupo, los otros dos grupos, presentan homofilias completamente diferentes, el grupo de bisexuales presenta heterofilia, y el grupo de travestis una homofilia mayor que la de gays pero no tan alta como la bisexuales.

Por lo tanto podemos decir que la red social de HSH está estructurada del tipo centro periferia, para el caso del grupo de bisexuales la estructura de la red social es bipartita.

Figura 21 - Homofilia, orientación sexual



Para el caso de la variable trabajo sexual, la homofilia en el grupo de trabajadores sexuales es de 0.24, lo que indica un tendencia al reclutamiento intra grupo, a diferencia de los no trabajadores sexuales, donde su homofilia es prácticamente cero, lo que indica que no hay preferencia en el reclutamiento. La red formada por trabajadores sexuales es de cohorte.

Figura 22 - Homofilia, trabajo sexual

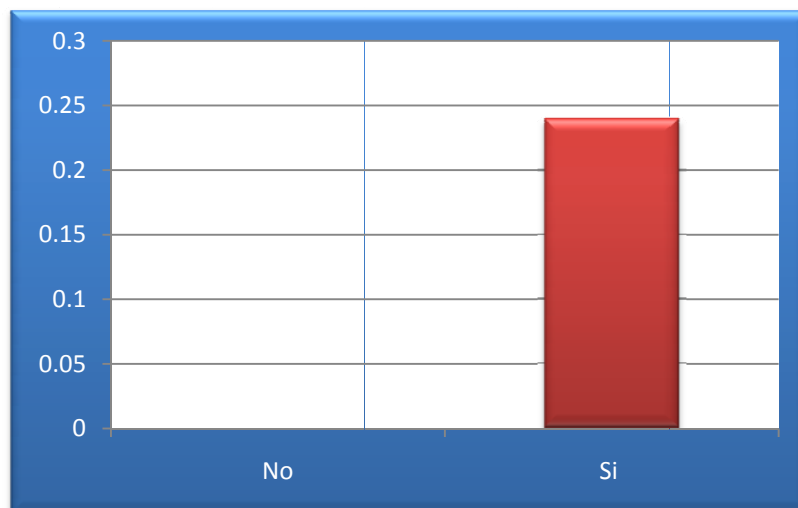
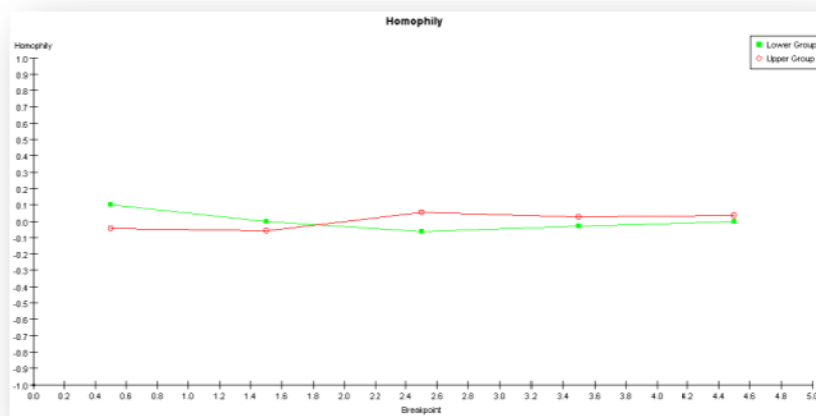


Figura 23 - Homofilia, edad



Fuente: Elaboración Propia

3.3.1.8. Estimación de la varianza

El RDS tiene la capacidad de producir estimaciones de varianza, a diferencia de otros métodos de muestreo de referencia en cadena. Las estimaciones de la varianza, usando el método bootstrapping, son fundamentalmente para establecer la confianza en una estimación. Para el presente caso RDS repite cinco mil veces la muestra, una estimación única de cada una de esas muestras es calculada y una distribución de estas estimaciones es determinada.

Independientemente del tamaño de la población y el tamaño de la muestra aleatoria, tomando repetidamente muestras aleatorias del mismo tamaño de la misma población y calculando las estimaciones de la varianza en cada muestra, dará agrupaciones alrededor del valor exacto de la varianza de la población⁶⁷.

Las varianzas intragrupo se pueden ver en las tablas 5 y 7, donde a partir del cálculo utilizando el método bootstrap para derivar estimaciones de la varianza⁶⁸, se obtiene para la variable de orientación sexual

	Gay	Bisexual	Travesti
Error estándar	0.033	0.035	0.022

⁶⁷ Rosner, 2000

⁶⁸ Volz, 2008; Salganik, 2006; Efron y Tibshirani, 1993

Y para la variable de trabajo sexual

	Trabajador sexual	No es trabajador sexual
Error estándar	0.023	0.023

Adicionalmente a esto necesitamos tener un estimador para una característica particular, en este caso tomamos aquellas personas con diferente orientación sexual que realizan el trabajo sexual.

Se tiene

	Gay	Bisexual	Travesti
Realizan trabajo sexual	0.129	0.147	0.731

Y sea,

$$Var(\hat{p}) = \frac{p - p^2}{n} + \frac{(p_A - p_B)^2 \beta_1}{2n(1 - \beta_1)} - \frac{(p_A - p_B)^2 (\beta_1 - \beta_1^{n+1})}{2n^2(1 - \beta_1)^2} \quad (63)$$

Una estimación de la varianza en un muestreo aleatorio simple asume que las muestras no están correlacionadas, dando sólo el primer término $(p - p^2)/n$, que no tiene en cuenta la posible estructura comunitaria en la población oculta. Por ejemplo, para $H = 0.129$, $p_A = 0.13$ y $p_B = 0.15$, la $Var(\hat{p})$ es aproximadamente 1,5 veces la varianza de las estimaciones de una muestra aleatoria simple⁶⁹. En consecuencia, los intervalos de confianza determinado por la variación real son $\sqrt{1.5} \approx 1.2$ veces más amplia. Dicho de otra manera, la estructura comunitaria en este ejemplo efectivamente reduce el tamaño de la muestra en un tercio las estimaciones RDS basadas en una muestra de 500 individuos tienen la misma varianza que las estimaciones basadas en una muestra aleatoria simple de 335.

Para el ejemplo de las variables en estudio, se tiene que del grupo de gays, los que realizan trabajo sexual son el 13%, del grupo de los identificados como bisexuales, es el 15% y del grupo de travestis, es el 73%. Se realiza el cálculo de acuerdo a la ecuación (63) y se obtiene que el valor estimado del error estándar para la característica de trabajo sexual en los diferentes grupos es de 0.018.

⁶⁹ Goel, Salganik. 2009.

3.3.1.9. Intervalos de confianza

RDSAT se predetermina a intervalos de confianza de 95%. Se realiza un grupo de remuestros para los intervalos de confianza, el cálculo genera las tablas a continuación para las variables en estudio.

Tabla 12 - Remuestros de los intervalos de confianza para P, orientación sexual

Remuestros	Gay (0.403)	Bisexual (0.506)	Travesti (0.092)
50	0.352 – 0.489	0.432 – 0.559	0.060 – 0.135
100	0.346 – 0.464	0.440 – 0.574	0.052 – 0.136
1000	0.344 – 0.472	0.435 – 0.568	0.054 – 0.139
2500	0.341 – 0.472	0.433 – 0.571	0.054 – 0.140
5000	0.342 – 0.472	0.433 – 0.567	0.055 – 0.140
10000	0.341 – 0.470	0.433 – 0.571	0.054 – 0.139

Tabla 13 - Remuestros de los intervalos de confianza para P, trabajo sexual

Remuestros	Trabajador Sexual (0.167)	No trabajador sexual (0.833)
50	0.128 – 0.227	0.773 – 0.872
100	0.122 – 0.223	0.779 – 0.878
1000	0.124 – 0.215	0.786 – 0.876
2500	0.124 – 0.216	0.785 – 0.876
5000	0.122 – 0.214	0.786 – 0.878
10000	0.125 – 0.214	0.786 – 0.875

A partir de esta información se elige el número de 5,000 remuestros para utilizar los intervalos de confianza generados por RDSAT.

El intervalo de confianza es un intervalo aleatorio ya que sus extremos dependen de la muestra escogida. Los intervalos aquí construidos tienen una probabilidad del 0,95 de capturar el valor de la media poblacional (μ). Es decir, que al extraer una muestra de la población, existe una probabilidad igual a $1 - \alpha$ de que el intervalo que se calcule realmente recoja el valor μ .

3.3.1.10. Sesgo asociado a las estimaciones de proporción de la muestra (S)

La interpretación más importante a tomar en cuenta es si las S son sesgadas. La salida de datos proporciona información sobre si la muestra tiene baja o sobre estimación de un determinado tipo, la homofilia es útil para explorar si existe reclutamiento preferencial. Los tamaños de red son útiles para explorar si un determinado tipo de persona tiene más “camino” que otros y por lo tanto más posibilidades de reclutamiento.

Como resumen podemos decir que; a) la proporción muestral (S) es el cálculo de dividir el número de entrevistados con la característica bajo estudio, por el total de la muestra. Sin embargo, la S no es representativa de la población, ya que se calcula sin tener en cuenta las medias muestrales de la red.

Por lo tanto, representa la estimación que se pueden encontrar en una muestra normal de referencia en cadena, b) estimadores de proporción de la población (P), es la proporción de estimaciones calculadas utilizando RDSAT descrito anteriormente. Si todos los supuestos y requisitos RDS se cumplen, la proporción poblacional debe ser representativa de las características que se encuentran en la población objetivo y c) el equilibrio muestral es una estimación de la proporción de la muestra de dos (o más) grupos distintos en el momento de convergencia. Por ejemplo, si la convergencia se alcanza por la tercera ola de reclutamiento (es decir, la composición de la muestra se mantiene estable), entonces el equilibrio es la proporción de subgrupos en el punto de convergencia. A diferencia de la proporción poblacional, el equilibrio deriva proporciones sin tener en cuenta las diferencias en los tamaños de la red. El equilibrio indica que la muestra ha alcanzado la convergencia, una indicación importante de que la muestra no está sesgada por la selección no aleatoria de las semillas.

La proporción muestral relacionada con el equilibrio indica que existe un sesgo mínimo introducido por la elección de semillas en la muestra y que las expectativas teóricas del proceso de reclutamiento de la muestra corresponden a un proceso de Markov.

Cuando la proporción muestral es similar a la proporción poblacional y cuando ambas estimaciones entran en los intervalos de confianza, hay un sesgo mínimo en la muestra. Sin embargo, la diferencia entre la S y la P va a ayudar a explicar los sesgos en la muestra.

Si la S para el grupo A es mayor que la P para el grupo A , podría indicar que existe una sobreestimación del grupo A en la muestra. La más adecuada estimación RDS para describir las características de la población de interés es la P . La siguiente sección describe con más detalle la forma de evaluar el sesgo cuando las diferencias entre los S y P son grandes.

3.3.1.11. Explorando las diferencias entre la muestra y las estimaciones de población

Hay varias formas para determinar las causas de las diferencias entre la muestra y la población estimada.

El tamaño medio de la red - Grandes diferencias en el tamaño medio de la red entre los grupos puede introducir un sesgo en la S . Uno de los grupos que tengan un tamaño medio de red más grande indica que ese grupo cuenta con más caminos en la red y por tanto, más posibilidades de ser reclutado. El uso de la media armónica de los tamaños de la red media debe minimizar el sesgo de las personas con tamaños más grandes de red (que reclutan a un ritmo más rápido). La metodología de muestreo RDS no puede controlar los sesgos relacionados con el error sistemático de los participantes al declarar con inexactitud el tamaño de sus redes, RDS funciona mejor cuando se produce una salida que tiene tamaños promedio de red igual en todos los grupos.

Patrones de reclutamiento - Grandes sistemas de reclutamiento diferencial puede explicar algunas de las diferencias entre la S y la P . Por ejemplo, si el grupo B recluta más frecuentemente a otros del grupo B , este grupo puede ser excesivamente representado en la muestra. Sin embargo, el reclutamiento

diferencial no debería sesgar a P ya que podemos ajustar el reclutamiento diferencial a través del ajuste demográfico. Patrones similares de reclutamiento hace que el rendimiento de los resultados sea más preciso.

Homofilia - Los altos niveles de homofilia pueden dar cuenta de algunas de las diferencias entre los S y P. Una homofilia de 0,8 en el grupo A indica un *agrupamiento* alto entre los miembros del grupo A. El grupo A el 80% del tiempo recluta a otros del grupo A y recluta al azar el resto del tiempo (20%). Una elevada homofilia 0.8 sesgaría la S porque los patrones de reclutamiento preferencial es la causa de que los del grupo A estén sobre-estimados.

Por otra parte, una baja homofilia de 0,03 para el grupo B indica que este grupo forma lazos dentro del grupo sólo el 3% del tiempo y recluta al azar el resto del tiempo.

Es aceptable si ambos grupos tienen homofilias iguales (es decir, ambos grupos tienen homofilia iguales) que anulan el sub o sobre muestreo debido a los lazos de reclutamiento del grupo de pertenencia⁷⁰. Sin embargo, lo mejor es no tener ni alta ni baja homofilia (cerca a 1 o -1), sino no tener homofilia u homofilia en torno a cero.

Falta de equilibrio - Es importante que la muestra alcance el equilibrio en las proporciones que se mide (en el que no cambian las características de la muestra, no importa cuantas más olas de reclutamiento se produzcan).

Si existe falta de equilibrio de ciertas variables, estas variables pueden ser omitidas en el análisis de rutina y decir que son resultados poco fiables.

Tabla 14 - Diferencias entre muestra y población, orientación sexual

	S	P	IC
Gay	0.556	0.403	0.342 – 0.496
Bisexual	0.341	0.506	0.433 – 0.567
Travesti	0.103	0.092	0.055 – 0.140

Tabla 15 - Diferencias entre muestra y población, trabajo sexual

	S	P	IC
Trabajador Sexual	0.169	0.167	0.122 – 0.214
No es Trabajador Sexual	0.831	0.833	0.786 – 0.878

Otra forma de calcular la diferencia media entre la *distribución muestral* y la *muestra en equilibrio*, se muestra en la línea 10 en la Tabla 5 y la línea 8 en la Tabla 7.

⁷⁰ Heckathorn, 2002a

El cálculo es bastante simple, para el caso de trabajo sexual es $|(Ea-SDa) + (Eb-SDb)/2|$, el resultado debe expresarse en positivo, dado pues lo que importa es el tamaño de la diferencia.

A esta diferencia media se denomina “tolerancia”⁷¹, la misma que si es menor de 2% indica una composición muy cercana entre la composición de la muestra real y la composición de la muestra en equilibrio teórica por lo que se puede tomar ese valor como una referencia válida.

En este caso la diferencia media es de 0.4% para orientación sexual y 0.9% para trabajo sexual, por lo que podemos concluir que ambas variables son cercanas entre la composición de la muestra y la composición de la muestra en equilibrio teórico.

⁷¹ Heckathorn (1997) y Wang et. al. (2005)

Capítulo 4: Conclusiones y Recomendaciones

4.1. Conclusiones

Las muestras de referencia de la cadena son un medio muy útil para recoger muestras de poblaciones ocultas. Se demuestra que cuando se maneja adecuadamente, las muestras de referencia en cadena pueden producir estimadores que son asintóticamente insesgados.

Se demuestra que este método de muestreo y sus estimaciones son fiables, aún cuando las semillas no son seleccionadas aleatoriamente, a pesar de que esto se ha demostrado que este método de referencia en cadena es confiable.

La muestra nos da información de las personas que componen la población y también de la red que los conecta. El método nos da una amplia cobertura entre la población, por la expansión geométrica del proceso de reclutamiento. La muestra obtenida a través de este procedimiento de muestreo ofrece una opción aplicable y estadísticamente válida para casos donde los informantes son de difícil acceso. Este tipo de muestreo ofrece el análisis de la estructura social en la que está conformada la red social, a través de sus miembros y las conexiones que tienen entre sí.

El limitar a través de cupones la sobre representación de alguno de los grupos, es una de las diferencias principales que presenta este método de muestreo, además de la introducción de un sistema de incentivos para los reclutados.

En primer lugar, para una adecuada aplicación de RDS el estudio debe estar precedido por un amplio trabajo de campo, que proporciona un conocimiento de la población bajo estudio.

En segundo lugar, el conocimiento en profundidad de la población proporciona la base para seleccionar las semillas y generar la confianza de los participantes en el proceso. En tercer lugar, se debe invertir tiempo y esfuerzo en los procesos de comunicación a los participantes la importancia de su participación en el desarrollo de la investigación.

Una de las principales desventajas de esta metodología es que la tasa de reclutamiento no se puede predecir, por lo tanto no sabemos cuánto tardará en ser recolectada toda la muestra, especialmente si no se tiene un conocimiento pleno del contexto de la población que se desea investigar, esto puede ocasionar que el tamaño de las redes no sea el esperado o se encuentren demasiadas redes pequeñas y encapsuladas de las cuales es muy difícil salir. Otra de las dificultades que se tiene es el tipo de incentivo, que si es inadecuado disminuirá el entusiasmo de los participantes.

Para el método RDS se ha realizado un trabajo considerable en las estimaciones puntuales, sin embargo otros parámetros de interés, como los coeficientes de correlación o coeficientes de regresión continúan sin desarrollarse.

En cuanto a la aplicación misma, se ha podido encontrar que existe una profundidad sociométrica, que permite el desarrollo de este tipo de metodologías, además se ha encontrado que el nivel de participación ha sido el esperado.

Las preguntas iniciales son preponderantes para la construcción de los tamaños de las redes, si existen muchas variables en el estudio, se deberá realizar una pregunta para cada una de estas variables y medir el tamaño de la red en cada una.

La variable orientación sexual presenta dos estructuras de red social, esto se debe tomar en consideración en futuras investigaciones.

La red de HSH está estructurada del tipo centro periferia, para el caso del grupo de bisexuales la estructura es bipartita.

La estructura de trabajo sexual era la esperada, con grado de homofilia relativamente alto, por la interrelación al interior del grupo y particularmente por el oficio que tienen, los otros grupos son más distantes.

Para ambas variables se logra el equilibrio entre la 3ª y 4ª ola, lo cual es un indicativo muy bueno, cuanto más tarde la muestra en entrar al equilibrio, más olas serían necesarias para poder encontrar los valores poblacionales, dado que no se consideraría representativa la muestra, el hecho de tener que aumentar el tamaño de la muestra o tener que ampliar por más olas tiene riesgos inherentes, el principal es la profundidad sociométrica, eso quiere decir que no existiría población para ampliar las olas y lo obtenido no sería suficiente para el cálculo de los estimadores.

Los factores de expansión dependen del pleno conocimiento del tamaño poblacional que hasta el momento en nuestro país es desconocido, actualmente solo se cuentan con estimaciones realizadas por ONUSIDA que afirman que 1 de cada 10 hombres ha tenido relaciones homosexuales sin detallar la caracterización de la preferencia sexual, por lo tanto, se puede dar como conclusión final que una vez se tenga los valores reales de las estimaciones poblacionales de hombres que tienen sexo con hombres, los estimadores encontrados en este estudio serán válidos y muy aproximados a los valores reales.

4.2. Recomendaciones

A partir de ese trabajo se ha visto la utilidad de la aplicación de este nuevo sistema de muestreo, en especial en este tipo de poblaciones que son estigmatizadas y que no darían información a través de otros sistemas de muestreo o el sesgo por el nivel de mentira en las respuestas invalidaría el estudio. Por lo tanto se ha demostrado que las características aplicadas permiten tener confiabilidad en la información recolectada.

Por lo anterior se debe considerar ampliar la aplicación de este sistema de muestreo a otro tipo de poblaciones, las mismas que cumplen con la característica básica de no tener marcos muestrales o los existentes son incompletos y por lo tanto no son nada útiles; por ejemplo se puede tomar en cuenta a comerciantes minoristas, que si bien existen registros estos no son accesibles y no son confiables. Otro ejemplo de este tipo de población, son los orfebres, cuya información puede ser muy bien utilizada por la banca privada para fomentar créditos.

Otra recomendación que surge de este trabajo, está referida a la capacidad de subvención de este tipo de estudios, los cuales son onerosos, tanto en dinero en efectivo como en insumos y tiempo. Por lo que se debe considerar estos factores al momento del diseño de cualquier estudio utilizando esta metodología.

Carta de conformidad de la institución



Asociación Proyecto de Salud Comunitaria
Organización No Gubernamental

La Paz, 10 de noviembre de 2010

Señor
Lic. Fernando Rivero
Director a.i. Carrera de Estadística
Universidad Mayor de San Andrés
Presente.-

Ref.- Informe final de trabajo dirigido Postulante Wilson René Alarcon Flores

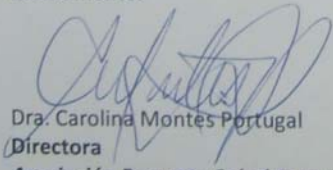
De mi consideración

Mediante la presente me permito hacer llegar a su Autoridad lo siguiente:

En el trabajo realizado en el periodo 01 de marzo al 15 de noviembre de 2010, por el señor Wilson René Alarcon Flores con CI 2477939 LP, universitario proponente para la licenciatura en Estadística, que con el cargo de investigador en el "Estudio de Conocimientos, Actitudes y Comportamientos asociadas al VIH/SIDA en población de Hombres que tienen Sexo con Hombres (HSH)" desempeñó su trabajo con eficiencia y responsabilidad, quedando la institución en plena conformidad con la elaboración y resultados del mencionado estudio.

Agradecida por la colaboración y deseándole éxitos en su vida profesional al proponente, es cuanto informo a su Autoridad.

Atentamente.


Dra. Carolina Montés Portugal
Directora
Asociación Proyecto Salud Comunitaria



cc. Archivo
Interesado

Calle Francisco de Miranda # 2239 – Miraflores
Telf. 2240570
La Paz, Bolivia

Bibliografía

Abdul-Quader, A. S., Heckathorn, D. D., Sabin, K. y Saidel, T. (2006). Implementation and Analysis of Respondent Driven Sampling: Lessons Learned from the Field. *Journal of Urban Health*.

Abu S. Abdul-Quader, Douglas D. Heckathorn, Courtney McKnight, Heidi Bramson, Chris Nemeth, Keith Sabin, Kathleen Gallagher, and Don C. Des Jarlais (2006). Effectiveness of Respondent-Driven Sampling for Recruiting Drug Users in New York City: Findings from a Pilot Study. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, Vol. 83, No. 3

Goel S., Salganik, M. J. (2009). Respondent-driven sampling as Markov chain Monte Carlo, *John Wiley & Sons, Ltd*

Heckathorn, D. (1997). Respondent Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*.

Heckathorn, D. (2002). Respondent Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Social Problems*.

Louise Clark, (2006). Manual para el mapeo de redes como una herramienta de diagnóstico. *Programa FIT-DFID, Bolivia al Proyecto FIT 3 (RedCampo)*.

Magnani R, Sabin K, Saidel T, Heckathorn D. (2005) Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*

Ramirez-Valles, J., Heckathorn, D.D., Va'zquez, R., Diaz, R.M., and Campbell, R.T. (2005). From Networks to Populations: The Development and Applications of Respondent-Driven Sampling Among IDUs and Latino Gay Men. *AIDS and Behavior*.

Salganik, M. J. (2006). Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling, *Journal of Urban Health*.

Salganik, M. J. y Heckathorn, D. (2004). Sampling and estimation in hidden populations using Respondent- Driven Sampling, *Sociological Methodology*.


Salaam Semaan, Jennifer Lauby and Jon Liebman (2002). Street and Network Sampling in Evaluation Studies of HIV Risk-Reduction Interventions. *AIDS Rev*.

Wejnert Cyprian, (2008). An Empirical Test Of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, And Out-Of-Equilibrium Data. *Forthcoming, Sociological Methodology*.

ONUSIDA (2000). Programas Nacionales de SIDA, Guía para el monitoreo y la evaluación. *ONUSIDA/00.17E (Original: inglés, junio de 2000)*.

Anexos

4.3. Boleta de encuesta

 <p>Asociación Proyecto de Salud Comunitaria</p>		“CONOCIMIENTOS, ACTITUDES Y COMPORTAMIENTOS RELACIONADOS AL VIH”									
		Población entrevistada: Hombres que tienen sexo con hombres									
Sección I: Para uso interno											
Boleta No											
Código RDS											
Código del entrevistador											
Fecha de la entrevista	/	/	Hora de la entrevista			:					
Esta entrevistas está dirigida a población de hombres que tienen sexo con hombres (es decir, gays, homosexuales, bisexuales, travestis, trans), por lo tanto quisiera saber ¿si Ud. ha tenido relaciones sexuales con otro hombre en los 12 últimos meses?						1. Sí		2. No			
										Si la respuesta es 2. NO suspenda la encuesta	
Sección II. Datos del reclutado											
1. Edad del entrevistado				En años cumplidos							
2. Nivel de instrucción	1. Primaria			2. Secundaria							
	3. Técnico (medio o superior)			4. Universitaria incompleta							
	5. Universitaria completa			6. Solo sabe leer y escribir							
3. Lugar de nacimiento	País					Ciudad					
4. Ocupación principal que le genera ingresos económicos	1. Trabajador dependiente			2. Trabajador independiente							
	3. Ingresos particulares			4. Trabajo sexual							
	5. Vive con sus padres			6. Su pareja lo mantiene							
	7. Otro			99. No responde							
5. Si Ud. tiene ingresos propios, ¿aproximadamente cuanto es su ingreso mensual?	1. No tienen ingresos propios			2. menos de 680B\$							
	3. de 681 a 1400 B\$			4. de 1401 a 3000 B\$							
	5. más de 3000 B\$			99. No responde							

6. ¿Tiene ud. personas dependientes?	1. Sí →	Cuántas personas	—	2. No
7. ¿Con quién vive actualmente?	1. Solo (a)			2. Con sus padres o familiares
	3. Con una pareja hombre, transexual o travesti			4. Con una pareja mujer
	5. Otro _____			99. No responde
Sección III. Uso de drogas y otras sustancias				
8. Durante las últimas 4 semanas ¿Con qué frecuencia ha ingerido bebidas alcohólicas?	1. Todos los días			2. Por lo menos una vez a la semana
	3. Más de una vez a la semana			4. Ni una vez
	99. No responde			
9. ¿En los últimos 12 meses consumió algún tipo de droga?	1. Sí		No Pasar a la pregunta 13	99. No responde Pasar a la pregunta 13
10. ¿En los últimos 12 meses consumió alguna droga inyectable, como ser cocaína o heroína?	1. Sí		No Pasar a la pregunta 13	99. No responde Pasar a la pregunta 13
11. ¿La última vez que se inyectó drogas, uso una jeringa desechable o esterilizó la que ya tenía?	1. Sí		2. No	99. No responde
12. ¿Comparte Ud. su jeringa con otras personas, o utiliza la jeringa de otra persona?	1. Sí, siempre			2. Sí, a veces
	3. Nunca			98. No sabe si alguien más usa su jeringa
	99. No responde			

Sección IV. Antecedentes Sexuales				
13. ¿En términos de orientación sexual, Ud. cómo se identifica? (El entrevistado debe identificarse en un grupo)	1. Gay u homosexual		2. Bisexual	
	3. Travesti		4. Transexual	
	5. Heterosexual			
14. ¿En los últimos 12 meses ha tenido relaciones sexuales con mujeres?	1. Sí	2. No	99. No responde	
15. ¿La última vez que tuvo relaciones sexuales con una mujer, uso condón?	1. Sí	2. No	99. No responde	
16. ¿Con que frecuencia usó condón con las mujeres que tuvo relaciones sexuales?	1. Siempre		2. Casi siempre	
	3. Algunas veces		4. Nunca	
	98. No sabe		99. No responde	
17. ¿Hace cuanto tiempo tuvo relaciones sexuales con un hombre, transexual o travesti?	1. Un día		2. Una semana	
	3. 2 semanas		4. Más de 2 semanas	
	5. Más de un mes		6. Más de 3 meses	
	7. Más de 6 meses pero menos de 1 año		Más de un año Reconfirmar esta respuesta y terminar la entrevista	
18. ¿La última vez que tuvo relaciones sexuales con una pareja masculina, uso condón?	Pasar a la pregunta 20		2. No	99. No responde
19. ¿Por qué no uso condón?	1. No tenía condones		2. Son muy caros	
	3. Porque mi pareja no quiso		4. Yo no quise	
	5. No me gustan los condones		6. No pensaron que fuera necesario	
	7. Creo que los condones no sirven		8. Otro (Especificar): _____	
	99. No responde			

20. ¿En sus relaciones sexuales con parejas estables hombres, con qué frecuencia utiliza condón?	1. Siempre	2. Casi siempre	
	3. Algunas veces	4. Nunca	
	98. No sabe	99. No responde	
21. ¿Durante los últimos 12 meses tuvo relaciones sexuales con parejas ocasionales hombres?	Sí	No Pasar a la pregunta 23	99. No responde Pasar a la pregunta 23
22. ¿Con que frecuencia utilizó condón en sus relaciones sexuales con sus parejas sexuales ocasionales hombres?	1. Siempre	2. Casi siempre	
	3. Algunas veces	4. Nunca	
	98. No sabe	99. No responde	
23. ¿Dónde encuentra a la mayoría de sus parejas sexuales ocasionales hombres?	1. Bar discoteca	2. Calles o parques	
	3. Centro comercial	4. Internet – Chat	
	5. Por teléfono – celular	6. A través de otras personas	
	7. Cines	8. Casa	
	9. Hotel	10. Baño público	
	99. No responde		
Sección V. Trabajo Sexual			
24. ¿En los últimos 12 meses ha recibido dinero a cambio de relaciones sexuales?	1. Sí	No Pasar a la pregunta 33	99. No responde Pasar a la pregunta 33
25. ¿Ud. ha tenido relaciones sexuales a cambio de dinero con...?	1. Solo hombres	2. Solo mujeres	
	3. Solo Transexuales-travestis	4. Hombres y mujeres	
	99. No responde		
26. ¿Dónde consigue sus clientes?	1. Bar discoteca	2. Calles o parques	
	3. Centro comercial	4. Internet – Chat	
	5. Por teléfono – celular	6. A través de otras personas	
	7. Cines	8. Casa de masajes	
	9. Hotel	10. Baño público	
	11. Anuncios de prensa	12. Otra (especificar).....	
	99. No responde		

27. ¿Cuántos clientes atiende normalmente por día?	Número de clientes			
28. ¿Cuánto cobra en promedio por tener relaciones sexuales con un cliente?	En B\$ por ocasión			
29. ¿En su última relación sexual con un hombre por dinero ha usado condón?	Sí Pasar a la pregunta 31	2. No	99. No responde	
30. ¿Por qué no uso condón?	1. No tenían		2. Son muy caros	
	3. Porque el cliente pago más		4. El cliente no quiso	
	5. Yo no quise		6. No me gustan los condones	
	7. No pensaron que fuera necesario		8. Creo que los condones no sirven	
	9. Otro (especificar) _____		99. No responde	
31. ¿Con qué frecuencia utiliza condón con sus clientes?	1. Siempre		2. Casi siempre	
	3. Algunas veces		4. Nunca	
	98. No sabe		99. No responde	
32. ¿Ha hablado con alguno de sus clientes acerca de VIH/SIDA o ITS?	1. Sí, con todos		2. Sí, con algunos	
	3. No, con ninguno		99. No responde	
Sección VI. Relación con trabajadores sexuales				
33. ¿Durante los últimos 12 meses Ud. pago dinero para tener relaciones sexuales?	1. Sí	No Pasar a la pregunta 37	99. No responde Pasar a la pregunta 37	
34. ¿Dónde encuentra a los trabajadores sexuales hombres, travestis o transexuales?	1. Bar discoteca		2. Calles o parques	
	3. Centro comercial		4. Internet – Chat	
	5. Por teléfono – celular		6. A través de otras personas	
	7. Cines		8. Casa de masajes	
	9. Hotel		10. Baño público	
	11. Anuncios de prensa		12. Otra (especificar) _____	
	99. No responde			

35. ¿La última vez que tuvo relaciones sexuales con un trabajador sexual hombre, utilizó condón?	1. Sí		2. No	99. No responde
	Pasar a la pregunta 37			
36. ¿Por qué no uso condón?	1. No tenían			2. Son muy caros
	3. Porque el cliente pago más			4. El cliente no quiso
	5. Yo no quise			6. No me gustan los condones
	7. No pensaron que fuera necesario			8. Creo que los condones no sirven
	9. Otro (especificar) _____			99. No responde
Sección VII. Condón				
37. ¿Sabe donde conseguir condones?	1. Sí		2. No	99. No responde
38. ¿Dónde consigue los condones habitualmente?	1. Supermercado			2. Farmacia
	3. Centro de Salud/Hospital			4. CDVIR
	5. ONG			6. Grupos Gay
	7. Night Club			8. Bar o Disco
	9. Hotel			10. Motel/ hospedaje
	11. Amigos			12. Educador(a)
	13. Otro (Especificar) _____			99. No responde
39. ¿En qué situaciones no usa el condón?	1. Cuando esta bebido			2. Cuando esta drogado
	3. Cuando tiene prácticas sexuales con su pareja estable			4. Cuando tiene relaciones con una pareja conocida
	5. Cuando tiene relaciones con una pareja casual o desconocida			6. Otro (especificar) _____
	7. Siempre usa condón			8. Nunca usa el condón
	99. No responde			

Sección VIII. ITS					
40. ¿Ha escuchado sobre las infecciones de transmisión sexual?		1. Sí		2. No	99. No responde
41. ¿Ha asistido alguna vez a un servicio médico para el diagnóstico de alguna ITS o VIH?		1. Sí		2. No	99. No responde
42. ¿En los 12 últimos meses, ha asistido a un servicio médico para el diagnóstico de una ITS o VIH?		1. Sí		2. No	99. No responde
43. ¿Durante los últimos 12 meses, ha tenido alguno de los siguientes síntomas?		1. Secreción por el pene		2. Dolor/ardor al orinar	
		3. Ulceras/lagas/granos en el pene o ano		4. Ganglios inguinales inflamados	
		5. Verruga/condiloma en el pene/ano		6. Picazón en genitales/ano	
		7. Otro (Especificar): _____		8. Ningún síntoma Pasar a la pregunta 46	
		99. No responde			
44. ¿Tomó medicamentos o algún tratamiento la última vez que tuvo una infección de transmisión sexual?		1. Sí		No Pasar a la pregunta 46	99. No responde Pasar a la pregunta 46
45. ¿Dónde consiguió los medicamentos?		1. Se la dio un amigo		2. En el CDVIR	
		3. En centro de Salud Público		4. En hospital o clínica privada	
		5. En la farmacia		6. En una ONG	
		7. En un grupo Gay		8. En el trabajo	
		9. Tomó medicina que tenía en la casa		10. Se la dio un medico tradicional	
		11. Otro (Especificar):		99. No responde	

Sección IX. Conocimientos acerca de VIH							
46. ¿Ha escuchado alguna vez sobre el VIH o el SIDA?		1. Sí		2. No		99. No responde	
47. ¿Ha participado en el último año en actividades de información o educación sobre VIH/SIDA?		1. Sí		No Pasar a la pregunta 49		99. No responde Pasar a la pregunta 49	
48. ¿Dónde?		1. Unidad de salud				2. ONG	
		3. Iglesias				4. En centros/organizaciones GLB	
		5. En organizaciones para PVVS				6. Otros (Especificar): _____	
		99. No responde					
49. ¿Puede una persona de aspecto saludable tener VIH?		1. Sí		2. No		98. No sabe	99. No responde
50. ¿Puede reducirse el riesgo de transmisión del VIH usando condones de manera correcta?		1. Sí		2. No		98. No sabe	99. No responde
51. ¿Puede reducirse el riesgo de transmisión del VIH manteniendo relaciones sexuales con una única pareja fiel y no infectada?		1. Sí		2. No		98. No sabe	99. No responde
52. ¿Se puede contraer el VIH por picaduras de mosquito?		1. Sí		2. No		98. No sabe	99. No responde
53. ¿Se puede contraer el VIH compartiendo comida con una persona infectada?		1. Sí		2. No		98. No sabe	99. No responde

54. ¿El VIH se puede transmitir por una aguja que había sido usada por alguien que ya estaba infectado?		1. Sí		2. No		98. No sabe		99. No responde
55. ¿EL VIH se puede transmitir por una mujer embarazada a su hijo?		1. Sí		2. No		98. No sabe		99. No responde
56. ¿Puede una mujer con VIH o SIDA transmitir el virus a su hijo/hija a través de la lactancia materna?		1. Sí		2. No		98. No sabe		99. No responde
57. ¿Es posible hacerse una prueba confidencial para saber si se está infectado con el VIH (el virus que causa el SIDA) donde reside actualmente?		1. Sí		2. No		98. No sabe		99. No responde
58. ¿Qué grado de riesgo de contraer VIH considera que tiene usted?		1. Alto				2. Medio		
		3. Bajo				4. Ninguno		
		99. No responde						

“MUCHAS GRACIAS POR SU TIEMPO Y TENGA LA SEGURIDAD QUE TODA LA INFORMACIÓN QUE DIO ES ESTRICTAMENTE CONFIDENCIAL”

4.4. Software RDSAT y manual (versión digital)

Se adjunta de manera digital el software RDSAT versión 5.6 de distribución libre para uso de investigación. Asimismo en el CD se incluye el manual de uso de esta versión.

Se puede hacer la descarga directa desde la página <http://www.respondentdrivensampling.org/>

Este software ha sido desarrollado por: Volz, E., Wejnert, C., Degani, I., and Heckathorn, D. D. 2007 Respondent-Driven Sampling Analysis Tool (RDSAT) Version 5.6. Ithaca, NY: Cornell University.