Adicciones
ISSN: 0214-4840
secretaria@adicciones.es
Sociedad Científica Española de Estudios sobre el Alcohol, el Alcoholismo y las otras Toxicomanías
España

Mantecón, Alejandro; Juan, Montse; Calafat, Amador; Becoña, Elisardo; Román, Encarna
Respondent-Driven Sampling: un nuevo método de muestreo para el estudio de poblaciones visibles y ocultas
Adicciones, vol. 20, núm. 2, 2008, pp. 161-169
Sociedad Científica Española de Estudios sobre el Alcohol, el Alcoholismo y las otras Toxicomanías
Palma de Mallorca, España

Disponible en: http://www.redalyc.org/articulo.oa?id=289122057008

Cómo citar el artículo
Número completo
Más información del artículo
Página de la revista en redalyc.org

Sistema de Información Científica
Red de Revistas Científicas de América Latina, el Caribe, España y Portugal
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

Respondent-Driven Sampling: un nuevo método de muestreo para el estudio de poblaciones visibles y ocultas

Alejandro Mantecón*, **; Montse Juan*; Amador Calafat*; Elisardo Becoña*, *; Encarna Román***

* Irefrea (Instituto Europeo de Estudios sobre la Prevención)
** Universidad de Alicante
***Universidad de Santiago de Compostela

Enviar correspondencia a:
Alejandro Mantecón. Correo e.: alejandro.mantecon@ua.es

Recibido: Septiembre de 2007
Aceptado: Febrero de 2008

RESUMEN

Este artículo presenta una variante de muestreo en cadena: el respondent-driven sampling (RDS). 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, el RDS se presenta como una mejora matemática del muestreo por bola de nieve orientada al estudio de poblaciones ocultas. Sin embargo, aquí tratamos de probar su validez para ser utilizado con aquellas poblaciones que no están registradas en marcos muestrales pero que, sin embargo, no ofrecen especiales dificultades para ser contactadas. En este trabajo explicamos el funcionamiento básico de RDS a partir de investigaciones sobre los jóvenes (de entre 14 y 25 años) que salen a divertirse con frecuencia los fines de semana, consumen alcohol y otras drogas y tienen relaciones sexuales. La investigación de campo se realizó entre mayo y julio de 2007 en Baleares, Galicia y Comunidad Valenciana. La presentación que se hace del estudio demuestra la utilidad de este tipo de muestreo cuando la población es accesible pero hay una dificultad que viene generada por la inexistencia de un marco muestral. No obstante, la muestra conseguida no es una muestra aleatoria representativa en términos estadísticos de la población objetivo. Ha de reconocerse que la muestra final obtenida es representativa de una "pseudo-población" que se aproxima, aunque dista de ser idéntica, a la población objetivo

Palabras clave: *respondent-driven sampling, técnicas de muestreo, consumo de alcohol, drogas recreativas.*

ABSTRACT

The paper introduces a variant of chain-referral sampling: respondent-driven sampling (RDS). This sampling method shows that methods based on network analysis can be combined with the statistical validity of standard probability sampling methods. In this sense, RDS appears to be a mathematical improvement of snowball sampling oriented to the study of hidden populations. However, we try to prove its validity with populations that are not within a sampling frame but can nonetheless be contacted without difficulty. The basics of RDS are explained through our research on young people (aged 14 to 25) who go clubbing, consume alcohol and other drugs, and have sex. Fieldwork was carried out between May and July 2007 in three Spanish regions: Baleares, Galicia and Comunidad Valenciana. The presentation of the study shows the utility of this type of sampling when the population is accessible but there is a difficulty deriving from the lack of a sampling frame. However, the sample obtained is not a random representative one in statistical terms of the target population. It must be acknowledged that the final sample is representative of a "pseudo-population" that approximates to the target population but is not identical to it.

Key words: *respondent-driven sampling, sampling technique, alcohol consumption, recreational drugs.*

INTRODUCCIÓN

El *Respondent-Driven Sampling* (RDS) es un método de muestreo diseñado originalmente para el estudio de poblaciones ocultas o de difícil acceso (Heckathorn, 1997). El propósito de este método es mejorar las carencias metodológicas de tipos de muestreo ya conocidos como la bola de nieve (*snowball sampling*) (Goodman, 1961), el muestreo a través de informantes clave (*key informant sampling*) (Deaux y Callaghan, 1985), o a partir de la utilización de mapas etnográficos (*targeted sampling*) (Watters y

Biernacki, 1989). El RDS recoge elementos de la teoría de redes y de los procesos de Markov para la elaboración de un procedimiento matemático que intenta paliar los problemas derivados de la falta de representatividad que limitan la validez de los resultados obtenidos por medio de los muestreos no probabilísticos o intencionales (Abdul-Quader, Heckathorn, Sabin y Saidel, 2006).

El artículo que se presenta no aspira a una explicación exhaustiva sino, más bien, a una invitación al conocimiento de este procedimiento. Se advierte que este método tiene apenas diez años de antigüedad y su desarrollo teórico-metodológico está más que activo. La depuración estadística que ha experimentado en los últimos años no se ha incluido. Aquí se expone su estructura metodológica esencial, que, por otro lado, ha dado pie a la publicación de un buen número de trabajos de investigación. La discusión en torno a la demostración de los supuestos matemáticos sobre los que se asienta no forma parte de los objetivos del trabajo, si bien, el lector interesado puede profundizar en la comprensión y crítica de los mismos sirviéndose de la bibliografía (la mayoría de los textos citados están publicados en internet, en este sentido se recomienda consultar en primer lugar la web: http://www.respondentdrivensampling.org/. Una crítica especialmente aguda puede hallarse en el artículo de Robert Heimer, 2005).

CARACTERÍSTICAS DE LA POBLACIÓN Y SELECCIÓN DE LOS ENTREVISTADOS

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 ellas tienen un especial recelo a ofrecer información a los investigadores ya que normalmente siguen comportamientos estigmatizados, mal vistos o ilegalizados (Heckathorn, 1997 y 2002). Para la aplicación del RDS, la población objeto de estudio debe presentar tres requisitos (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 sociométrica. No obstante, esa "profundidad" tiene unos límites que, a su vez, generan limitaciones inherentes a los métodos de muestreo realizados por encadenamientos 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.

Como en otros métodos similares, la aplicación del RDS se inicia con la identificación de unos informantes iniciales que cumplen la función de "semillas". Sin embargo, tres cuestiones lo hacen diferente (Heckathorn, 1997; Wang, Carlson, Falck, Siegal, Rahman y Li, 2005):

1) las semillas no son seleccionadas aleatoriamente de la población objetivo con el fin de que identifiquen a pares que posteriormente el investigador abordará y entrevistará. Aquí, a las semillas se les solicita que directamente seleccionen a personas de la población objetivo (facilitando 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 los encadenamientos necesarios, se logra una estabilidad en la presencia porcentual de una serie de categorías grupales que se consideren significativas (por razón de sexo, grupos étnicos, grupos de edad, etc.). Estas categorías deben ser mutuamente excluyentes, es decir, si se pertenece a la categoría "negro" no es posible aparecer en la categoría "blanco" o en la categoría "latino", si se pertenece a la categoría de "afectados" por una determinada enfermedad no se puede formar parte al mismo tiempo de la de los "no afectados", etc. Se asume que la estabilidad de la muestra resulta independiente de la presencia relativa de las categorías de personas con que se inició el proceso de reclutamiento (Abdul-Quader et. al., 2006). 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, se sugiere 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 suelen 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 su artículo de 1997 Heckathorn advierte que, en situaciones de limitaciones presupuestarias en las que no se puede cumplir totalmente la doble recompensa, resulta más rentable concentrar los esfuerzos en el incentivo secundario). No obstante, se precisa aquí que la decisión de catalogar como oculta a una determinada población a veces se basa en criterios subjetivos del investigador. De tal modo, debe reconocerse una gradación de grises muy amplia entre una población listada en un registro de carácter público y una población marginal claramente estigmatizada. Por lo tanto, se acuerda que existe un criterio bastante flexible a la hora de decidir si es necesario emplear este sistema de recompensas. En algunos casos será imprescindible mientras que en otros puede que no haga falta; y

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

LA ELABORACIÓN DEL INTERVALO DE CONFIANZA

La lógica operativa de este muestreo no realiza estimaciones sobre la población estudiada directamente a partir de la muestra seleccionada. En cambio, los datos recogidos en la muestra son utilizados para estimar valores sobre una red social producida a través de la dinámica de encadenamientos obtenida en el contacto y selección de unos informantes por otros que, a modo de paso intermedio, es la que servirá para hacer las estimaciones poblacionales. Para poder llevar a la práctica el muestreo estas estimaciones deben ir acompañadas de unos intervalos de confianza. El intervalo de confianza se obtiene tras una serie de pasos (Salganik, 2006):

1) a partir de la muestra original obtenida se generan una serie de muestras réplica, creadas a través de un procedimiento de auto-reposición (*bootstrap procedure*). Se trata de llevar a cabo un muestreo aleatorio con reposición tomando la muestra original. La muestra original se divide en dos grupos: los seleccionados por personas del grupo A (A_{sel}) y los seleccionados por personas del grupo B (B_{sel}). A_{sel} podrían ser los individuos de la muestra original seleccionados por mujeres. Nótese que aquí estarían incluidos tanto hombres como mujeres. La muestra réplica se empieza a elaborar tras seleccionar aleatoriamente a una "semilla" de la muestra original. Seguidamente, tras comprobar el grupo al que pertenece la "semilla", se extrae con reposición a alguien A_{sel} o B_{sel} según corresponda. Por ejemplo, si la semilla seleccionada aleatoriamente para iniciar la muestra réplica es una mujer, a continuación habrá que elegir aleatoriamente algún miembro del grupo de la muestra que está integrado por todos los hombres y mujeres que fueron reclutados por mujeres (el segundo grupo sería el de los hombres y mujeres que fueron reclutados por hombres). Después, se comprueba si esta nueva persona seleccionada es un hombre o una mujer y se vuelve a repetir la operación. Si en este caso el seleccionado fuera un hombre a continuación habrá que elegir aleatoriamente a una persona (hombre o mujer) que hubiera sido seleccionada por un hombre (es decir, a una persona del segundo grupo). Este proceso continúa hasta que la muestra réplica alcanza el mismo tamaño que la muestra original. El número de muestras réplicas que hay que producir es elevado por lo que se hace imprescindible la asistencia de un programa informático (en la web citada en la introducción se puede acceder al software específico *Respondent Driven Sampling Analysis Tool - RDSAT*);

2) en el segundo paso hay que calcular la estimación de la proporción en la población de cada una de las categorías grupales para cada una de las muestras réplicas generadas (en el siguiente apartado se explica este asunto);

3) en el tercer paso las estimaciones obtenidas de todas las muestras réplica son utilizadas para construir un intervalo de confianza. Así, un intervalo de confianza al 95% basado en una aproximación normal sería: $[\hat{P}_A - 1{,}96\ \hat{se}(\hat{P}_A),\ \hat{P}_A + 1{,}96\ \hat{se}(\hat{P}_A)]$, donde el error estándar estimado, $\hat{se}(\hat{P}_A)$, es la desviación estándar de las estimaciones obtenidas en las muestras réplica.

Todo este procedimiento se entiende mejor a partir del ejemplo de un estudio concreto.

ILUSTRACIÓN A TRAVÉS DE UN ESTUDIO

A continuación se expone la práctica de este método con un estudio realizado desde Socidrogalcohol en Baleares, Galicia y la Comunidad Valenciana, entre mayo y julio de 2007, que ha dado ya lugar a distintas publicaciones (Calafat, Adrover, Blay y Juan, 2008; Calafat, Juan, Becoña, Mantecón y Ramón 2008).

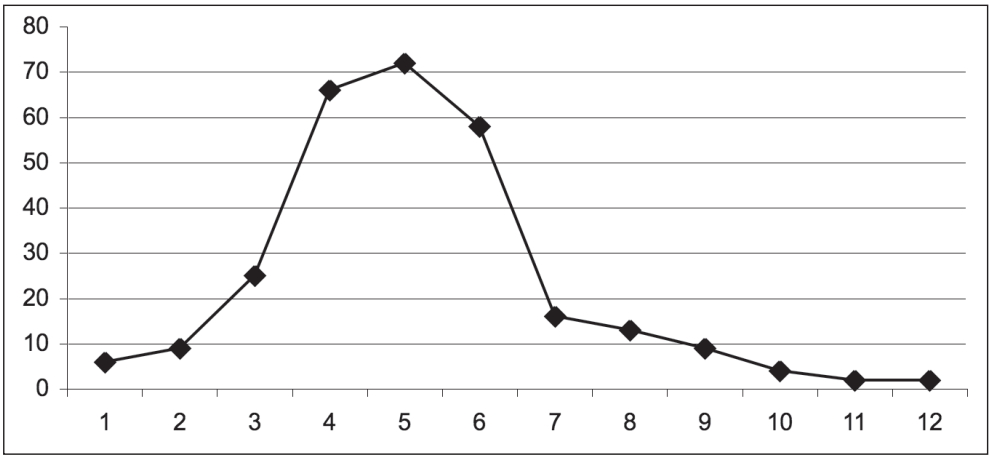

Figura 1. Entrevistados en cada una de las olas o cadenas

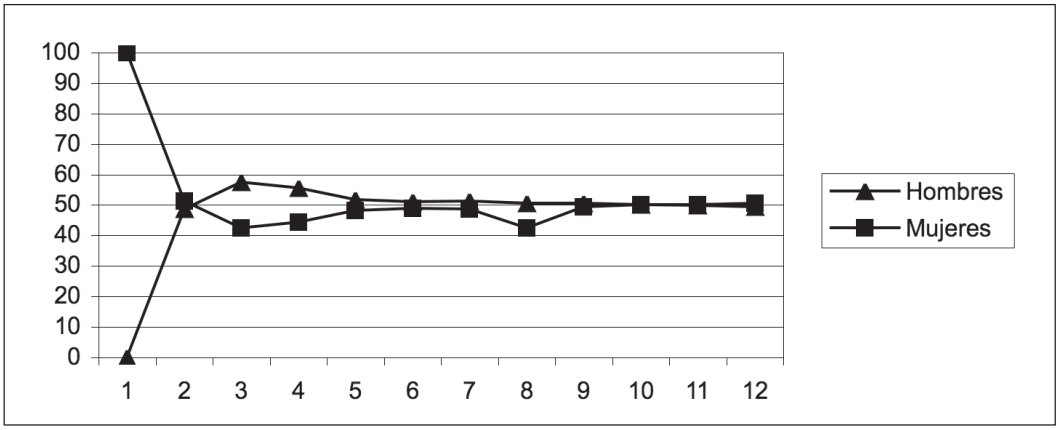

Figura 2. Evolución del % de entrevistados por género en cada una de las olas o cadenas

La población objetivo estaba integrada por jóvenes de entre 14 y 25 años, asiduos a salir de marcha, que se reconocían como consumidores de alcohol y de alguna otra droga y que habían tenido relaciones sexuales en los últimos meses. Esta población no se muestra "muy oculta" (aunque sí que resultaba difícil de abordar ya que las preguntas que se formulaban en el cuestionario de la encuesta podían dar lugar a respuestas socialmente sancionadas que, en última instancia, retraían considerablemente la participación de los potenciales entrevistados). En este sentido, se apunta que uno de los objetivos de la investigación era explorar la viabilidad de la aplicación del RDS para el estudio de poblaciones en las que la principal dificultad para el acceso a las mismas fuera la ausencia de marcos muestrales y no tanto la disponibilidad de los individuos para ser entrevistados.

El trabajo de campo dio lugar a doce olas de reclutamiento, en las que se entrevistó a 282 informantes. El número de entrevistados conseguidos en cada una de las olas se refleja en la Figura 1. El número de olas recomendables es variable. Si todo va bien, la muestra tiende a estabilizarse definitivamente entre la sexta y la séptima ola. Aunque se aprecie una estabilidad de las categorías grupales con un número menor de olas se aconseja alcanzar doce o más para asegurar que la muestra obtenida se aproxima realmente a las características de la población objetivo (Heckathorn, 2002).

Al ser uno de los objetivos del estudio la comprensión del funcionamiento del RDS, se utilizaron sólo dos categorías muy sencillas: el género (masculino y femenino) y los grupos de edad (14-18 y 19-25). En las Figuras 2 y 3 se aprecia la dinámica de reclutamiento en cada una de las categorías y la consecución de su estabilidad.

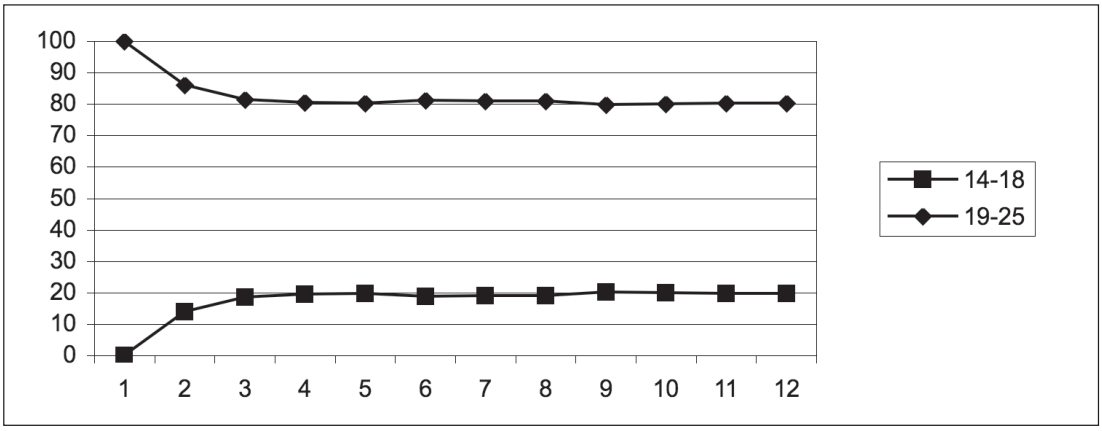

Figura 3. Evolución del % de entrevistados por grupos de edad en cada una de las olas o cadenas

Como se observa en los gráficos, en la primera ola obtenida a partir de las "semillas" únicamente se entrevistaron a seis personas, en concreto a seis mujeres de entre 19 y 25 años. La posterior dinámica de encadenamientos acaba por ofrecer, tras la duodécima ola, un 20,5% de entrevistados de entre 14 y 18 años y un 48,9% de hombres, a pesar de no haber ninguno de ambos grupos en la ola inicial. Aunque, tal y como puede verse, la lógica de la selección a través de las olas resuelve en gran medida este problema, también es verdad que cuanto más heterogénea sea la ola inicial la estabilidad tiende a alcanzarse antes.

En las Tablas 1 y 2 se presentan los datos básicos y, seguidamente, se detalla cómo se han elaborado y la interpretación de los mismos.

Tabla 1. Recuento por género

	Género del reclutado		
Género del reclutador	**Femenino**	**Masculino**	**Total**
Femenino			
Recuento de reclutados	94	46	140
Proporción muestral, S	0,671	0,328	1
Masculino			
Recuento de reclutados	44	98	142
Proporción muestral, S	0,309	0,690	1
Distribución total de los reclutados	138	144	282
Distribución muestral, SD	0,489	0,510	1
Muestra en equilibrio, E	0,485	0,514	
Diferencia media entre SD y E		0,4%	
Tamaño medio de la red ajustado, N	17,583	19,058	
Estimación de la proporción poblacional, P	0,505	0,494	
Homofilia, H	0,336	0,387	

Tabla 2. Recuento por grupo de edad

		Edad del reclutado		
Edad del reclutador		**14-18**	**19-25**	**Total**
14-18				
Recuento de reclutados	(1)	32	22	54
Proporción muestral, S	(2)	0,592	0,407	1
19-25				
Recuento de reclutados	(3)	26	202	228
Proporción muestral, S	(4)	0,114	0,885	1
Distribución total de los reclutados	(5)	58	224	282
Distribución muestral, SD	(6)	0,205	0,794	1
Muestra en equilibrio, E	(7)	0,218	0,781	
Diferencia media entre SD y E	(8)		1,3%	
Tamaño medio de la red ajustado, N	(9)	26,39	17,075	
Estimación de la proporción poblacional, P	(10)	0,153	0,846	
Homofilia, H	(11)	0,519	0,255	

Se ilustra la explicación tomando como ejemplo la Tabla 2. Se han numerado las filas que recogen información para leerlas y explicarlas una a una sin pérdida alguna:

(1) *Recuento de reclutados*: la primera columna indica que 32 entrevistados de entre 14 y 18 años han sido seleccionados por entrevistados pertenecientes a su mismo grupo de edad. La segunda columna indica que 22 entrevistados de entre 19 y 25 años han sido reclutados por entrevistados de entre 14 y 18 años. La tercera columna es la suma de las dos anteriores e indica que 54 es el total de entrevistados seleccionados por entrevistados de entre 14 y 18 años. Conocer las características del reclutador de cada entrevistado es posible gracias al registro de los cupones (elaborados por los investigadores) que los entrevistados se pasan entre sí cuando logran un reclutado nuevo (y que el nuevo entrevistado presenta al investigador de campo cuando tiene lugar la entrevista cara a cara) o, también, mediante el registro en el cuestionario de un código que permita conocer en qué ola de reclutamiento fue seleccionado cada individuo en concreto y tanto sus características grupales como las de quien lo seleccionó (a partir de ese código el programa informático tiene que saber a qué categorías pertenece el entrevistado y a qué categorías grupales pertenecía el entrevistado que lo reclutó). Por lo demás, esta fila se limita a presentar el recuento básico.

(2) *Proporción muestral*: la primera columna indica que 0,592 es la proporción de 32 sobre 54 (datos de la fila 1). La segunda columna indica que 0,407 es la proporción de 22 sobre 54. La tercera columna es la suma de las dos anteriores y siempre debe dar 1.

(3) y (4) En estas filas se vuelve a hacer lo mismo que en las dos anteriores, en este caso con los entrevistados seleccionados por jóvenes de entre 19 y 25 años.

(5) *Distribución total de los reclutados*: la primera columna es la suma de los datos recogidos en la primera columna de la fila 1 y en la primera columna de la fila 3, e indica que en la muestra hay un total de 58 entrevistados (32+26) de entre 14 y 18 años. La segunda columna es la suma de los datos recogidos en la segunda columna de la fila 1 y en la segunda columna de la fila 3, e indica que en la muestra hay un total de 224 entrevistados (22+202) de entre 19 y 25 años. La tercera columna indica el tamaño de la muestra obtenida, que en este caso es 282.

(6) *Distribución muestral*: la primera columna indica la proporción de 58 sobre 282, es decir la proporción de jóvenes entrevistados de entre 14 y 18 años sobre el conjunto de jóvenes que forman la muestra. La segunda columna indica esa misma proporción pero con los 224 jóvenes de entre 19 y 25 años.

(7) *Muestra en equilibrio*: se obtiene tras resolver un sencillo sistema de ecuaciones lineales en el que utilizamos los valores de las filas 2 y 4 para elaborar una matriz de transición de probabilidades, es decir:

$1 = E_a + E_b$

$E_a = S_{aa} E_a + S_{ba} E_b$

donde:

- E_a es el valor para la muestra en equilibrio de a (en este caso la composición muestral de los jóvenes de entre 14 y 18 años).
- E_b es el valor para la muestra en equilibrio de b (en este caso la composición muestral de los jóvenes de entre 19 y 25 años).
- S_{aa} es la probabilidad de que un entrevistado del tipo a haya sido reclutado por otro de su mismo grupo.
- S_{ba} es la probabilidad de que un entrevistado del tipo a haya sido reclutado por otro del tipo b.

$E_a = 0.592E_a + 0,114E_b$

$E_b = 1 – E_a$

$E_a = 0,592E_a + 0,114\ (1 – E_a)$

$E_a = 0,592E_a + 0,114 – 0,114E_a$

$0,522E_a = 0,114$

$E_a = 0,114 / 0,522 = 0,218$

$E_b = 1 – 0,218 = 0,781$

0,218 y 0,781 son los valores de la muestra en equilibrio para la categoría grupal de grupos de edad (fila 7).

En los trabajos de Heckathorn (1997 y 2002) se expone el modo de operar cuando las ecuaciones lineales tienen tres o cuatro incógnitas. Por ejemplo, si tenemos una categoría relativa a grupos étnicos en la que aparecen cuatro grupos distintos (blancos, latinos, negros y otros) se resolvería sin problemas un sistema de ecuaciones lineales con cuatro ecuaciones y cuatro incógnitas. No obstante, a la hora de calcular la estimación poblacional se producen algunas variaciones, por lo que se aconseja consultar antes los artículos mencionados. Por otro lado, se advierte que todo este procedimiento puede resolverse con el programa RDSAT antes apuntado. Aunque todas estas operaciones podrían hacerse con

papel, lápiz y calculadora será inevitable utilizar el programa para la posterior construcción de los intervalos de confianza y, una vez que hay que utilizar el programa para este fin, ya de paso ofrece todos los demás datos.

(8) *Diferencia media entre la Distribución muestral y la Muestra en equilibrio*:

(0,218 – 0,205) + (0,781 – 0,794) / 2 = 0,013

Los valores de las diferencias han de expresarse siempre en positivo, pues lo que importa es el tamaño de la diferencia. La diferencia media entre la muestra real y la muestra en equilibrio se denomina "tolerancia". Heckathorn (1997) y Wang et. al. (2005) explican que una tolerancia del 2% o menor indica una aproximación 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 del 1,3%.

(9) *Tamaño medio de la red ajustado*: en el cuestionario de la encuesta es imprescindible formular una pregunta relativa al número de personas de la población objetivo que conoce el entrevistado. Al diseñar el cuestionario el investigador ha de tener claro cuáles son las categorías grupales con las que va a aplicar el RDS, pues a la hora de redactar la pregunta se ha de precisar a cuántas personas conoce el entrevistado de cada una de las categorías grupales. Por ejemplo, si la población objetivo estuviera formada por personas afectadas por el VIH que consumen con asiduidad una determinada droga y las categorías grupales son tres grupos de edad y tres grupos étnicos debe recogerse a través de la pregunta del cuestionario a cuántas personas de la población objetivo conoce el entrevistado en cada una de las seis categorías.

Siguiendo con el trabajo que se utiliza como ejemplo, la media que resulta del cómputo de los 282 entrevistados es la siguiente: los jóvenes de entre 14 y 18 años conocen por término medio a 51,893 personas de la población objetivo y los jóvenes de entre 19 y 25 años a 50,371. Ahora bien, se ha considerado oportuno ajustar estas medias eliminando casos atípicos particularmente extremos que, se ha pensado, distorsionan la realidad. De tal modo, se decidió fijar como límite las 400 personas conocidas por cada entrevistado y se obvió a un pequeño grupo que afirmaba conocer a más de esas 400 personas. Así, el tamaño medio de la red ajustado para los jóvenes de entre 14 y 18 años es de 26,390 personas y el de los jóvenes de entre 19 y 25 años de 17,075 personas.

(10) *Estimación de los porcentajes poblacionales*: para hallar esta estimación se precisa del tamaño medio de la red (fila 9) y, de nuevo, de los valores de las filas 2 y 4 para elaborar una matriz de transición de probabilidades

$$P_a = S_{ba} N_b / S_{ba} N_b + S_{ab} N_a$$

donde:

- S_{ab} es la probabilidad de que un entrevistado del tipo b haya sido reclutado por otro del tipo a.
- S_{ba} es la probabilidad de que un entrevistado del tipo a haya sido reclutado por otro del tipo b.
- N_a = tamaño medio de la red para los entrevistados del grupo a.
- N_b = tamaño medio de la red para los entrevistados del grupo b.

P_a = (0,114) (17,075) / (0,114) (17,075) + (0,407) (26,39)

P_a = 1,947 / 1,947 + 10,74 = 0,153

$P_b = 1 – P_a$ = 0,846

(11) *Homofilia*: la homofilia indica la tendencia hacia el reclutamiento dentro del grupo de pertenencia y la heterofilia indica la tendencia hacia el reclutamiento fuera del grupo de pertenencia. El índice de la homofilia describe la extensión de los vínculos existentes dentro de un grupo. Un índice de H = 1 reflejaría una homofilia perfecta, es decir, todos los vínculos se han establecido con individuos que pertenecen al mismo grupo. Por el contrario, un índice de H = – 1 reflejaría una heterofilia perfecta, todos los vínculos se han establecido fuera del grupo de pertenencia (Heckathorn, 1997; Stormer et. al., 2006). Los índices altos de homofilia con frecuencia aparecen al evaluar grupos que pertenecen a una misma población pero que presentan una distancia de estatus muy acusada. El grupo ideal es aquel que está lo suficientemente estructurado como para garantizar una cierta profundidad sociométrica sustentada en un conocimiento mutuo entre sus miembros y que, al mismo tiempo, está lo suficientemente desestructurado como para que el reclutamiento de sus miembros simplemente refleje la presencia del grupo en la población objetivo: cada reclutado, al margen de su adscripción grupal, seleccionaría a una mezcla aleatoria de personas de la población objetivo independiente de su pertenencia grupal. Así, la probabilidad de que un informante seleccionase a otro de un determinado grupo sería igual a la proporción de personas de ese otro grupo en el total de la población objetivo, y no aparecerían problemas de sobre o infra-representación. En tanto que este criterio no se cumple,

en una población objetivo formada por grupos en los que la afiliación grupal afecta a la selección de nuevos informantes, las personas seleccionadas reflejan y reproducen un sesgo bajo el que subyacen factores de tipo cultural y situacional. Ese sesgo es evaluado por medio del índice de homofilia:

$$S_{aa} = H_a + (1 - H_a)\, P_a$$
$$S_{ba} = (1 - H_b)\, P_a$$

$$0{,}592 = H_a + (1 - H_a)\, 0{,}153$$
$$0{,}592 = H_a + 0{,}153 - 0{,}153H_a$$
$$0{,}439 = 0{,}847H_a$$
$$H_a = 0{,}439 / 0{,}847 = 0{,}518$$

$$0{,}114 = (1 - H_b)\, 0{,}153$$
$$0{,}114 = 0{,}153 - 0{,}153H_b$$
$$-\, 0{,}039 = -\, 0{,}153H_b$$
$$H_b = -\, 0{,}039 / -\, 0{,}153 = 0{,}255$$

En el caso de este estudio se aprecia que la homofilia del grupo de jóvenes de entre 14 y 18 años es de 0,518. Es decir, la afiliación grupal de los jóvenes de entre 14 y 18 años a su grupo de edad determina la selección de nuevos informantes el 51,8% de las veces, o lo que es lo mismo, solamente el 48,2% (100 – 51,8) de las nuevas selecciones se producen de manera aleatoria. En el grupo de jóvenes de entre 19 y 25 años el sesgo apuntado por el índice de homofilia interviene en menor medida, en concreto el 25,5% de las ocasiones. Si se lee la Tabla 1, puede apreciarse que en el caso de distinción grupal por razón de género los resultados no son mejores. Los motivos por los que este sesgo aparece aquí de una manera tan evidente pueden ser de diversa índole. En el caso de la investigación llevada a cabo es muy probable que la muestra con la que se ha trabajado esté integrada por subgrupos excesivamente "encapsulados" de acuerdo a las categorías grupales planteadas, lo que ha motivado un problema de selección endogrupal que, en definitiva, advierte de la prudencia con la que deben interpretarse los resultados de la encuesta.

El intervalo de confianza: como se apuntaba en el apartado anterior, para la construcción del intervalo de confianza es necesario calcular la estimación de la proporción en la población de cada una de las categorías grupales para cada una de las muestras réplicas generadas. En este estudio se trabajó a partir de la producción con el RDSAT de 2500 muestras réplica. El error estándar estimado es la desviación estándar de las estimaciones obtenidas en las muestras réplica. De tal modo, continuando con el ejemplo de la Tabla 2, para la categoría "grupos de edad" el programa elaboró los siguientes intervalos con un nivel de confianza del 95% ($\alpha = 0{,}05$):

Tabla 3. Intervalo de confianza

	Estimación de la proporción poblacional, P	Límite inferior	Límite superior
14-18	0,153	0,098	0,233
19-25	0,846	0,766	0,901

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 .

A MODO DE CONCLUSIÓN

En este artículo se presenta la estructura básica de un método de muestreo relativamente reciente diseñado para el estudio de poblaciones ocultas o de difícil acceso. Esta metodología de muestreo intenta paliar, al menos en cierta medida, algunas de las limitaciones que caracterizan a los métodos tradicionales de muestreo intencionales basados en el encadenamiento de entrevistados a través de los contactos establecidos por medio de las redes de una serie de informantes iniciales. No es el objetivo del artículo hacer una exposición exhaustiva sobre el tema, sino invitar al lector interesado a dar un primer paso hacia el conocimiento de un procedimiento que puede resultar de interés para aquellos investigadores dedicados al estudio de poblaciones "marginales". Aquí, además, se han explorado las posibilidades del RDS para el muestreo de poblaciones no marginales, parcialmente ocultas o no ocultas, cuyo impedimento para ser abordadas mediante un tipo estricto de muestreo probabilístico se halla sobre todo en la inexistencia de un marco muestral, pero no tanto en la dificultad de acceso a la población.

La muestra obtenida a través del procedimiento de selección de entrevistados que propone el RDS ofrece un punto de partida aceptable si se tienen en cuenta tanto los inconvenientes motivados por la dificultad en el acceso a los informantes como los problemas que surgen a la hora de evaluar la representatividad de los resultados que se obtienen tras el análisis. Esta muestra asume la existencia de una población que puede abordarse en los términos de una organiza-

ción, es decir, que ofrece una estructura de red social: sus miembros están conectados entre sí, por eso es imprescindible introducir en el cuestionario preguntas relativas a las características de la red del entrevistado. El procedimiento también introduce algunos elementos correctores muy provechosos, como es el caso de la limitación del número de reclutados por cada entrevistado (para evitar problemas de sobrerepresentación de las redes de una persona en concreto), y pone en práctica un sistema de incentivos que hace especial hincapié en la recompensa por la selección de nuevos informantes, susceptible de ser empleado de una u otra manera, o de prescindir de su uso, en función de las características específicas de la población que se estudie.

No obstante, la muestra conseguida no es una muestra aleatoria representativa en términos estadísticos de la población objetivo. Ha de reconocerse que la muestra final obtenida es representativa de una "pseudo-población" (Wang et. al., 2006) que se aproxima, aunque dista de ser idéntica, a la población objetivo. La representatividad de la muestra conseguida con el RDS puede examinarse en primer lugar a partir de la propia información que genera el procedimiento, esto es, comparando las proporciones muestrales de la muestra real con las composiciones de las proporciones poblacionales estimadas.

AGRADECIMIENTOS

Este estudio se ha realizado en parte con financiación de la Delegación del Gobierno para el Plan Nacional sobre Drogas según la Orden de convocatoria SCO/269/2007.

REFERENCIAS

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, 83,* 1-5.

Calafat, A., Adrover, D., Blay, N. y Juan, M. (2008). Relación del consumo de alcohol y drogas con la siniestralidad vial de los jóvenes españoles durante la vida recreativa nocturna. *Revista Española de Salud Pública.* Manuscrito remitido para publicación.

Calafat, A., Juan, M., Becoña, E., Mantecón, A. y Ramón, A. (2008). Sexualidad de riesgo y consumo de drogas en el contexto recreativo. Una perspectiva de género. *Psicothema* (En prensa).

Deaux, E. y Callaghan, J. W. (1985). Key informant versus self-report estimates of health behavior, *Evaluation Review, 9,* 365-368.

Goodman, L. A. (1961). Snowball Sampling, *Annals of Mathematical Statistics*, *32*: 148-170.

Heckathorn, D. (1997). Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations, *Social Problems, 44:* 174-199.

Heckathorn, D. (2002). Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations, *Social Problems*, *49*: 11-34.

Heckathorn, D. (2007). Extensions of Respondent-Driven Sampling: analyzing continuous variables and controlling for differential recruitment, *Sociological Methodology*, *37:* 151-207.

Heckathorn, D. y Jeffri, J. (2005). Assessing the Feasibility of Respondent-Driven Sampling: Aging Artist in New York City. Documento de trabajo. Research Center for Arts and Culture (Teachers College, Columbia University) editado en la web http://www.tc.columbia.edu/centers/rcac/pdf/FeasRep_12.pdf [consultado el 17 de junio de 2007].

Heimer, R. (2005). Critical Issues and Further Questions about Respondent-Driven Sampling: Comment on Ramirez-Valles, et al. (2005). *AIDS and Behavior*, *9*: 403-408.

Salganik, M. J. (2006). Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling", *Journal of Urban Health*, *83*: 98-112.

Salganik, M. J. y Heckathorn, D. (2004). Sampling and estimation in hidden populations using Respondent-Driven Sampling, *Sociological Methodology*, *34*: 193-239.

Stormer, A., Tun, W., Guli, L., Harxhi, A., Bodanovskaia, Z., Yakovleva, A., Rusakova, M., Levina, O., Bani, R., Rjepaj, K. y Bino, S. (2006). An Analysis of Respondent Driven Sampling with Injection Drug Users (IDU) in Albania and the Russian Federation. *Journal of Urban Health*, *83:* 73-82.

Wang, J., Carlson, R. G., Falck, R. S., Siegal, H. A., Rahman, A. y Li, L. (2005) Respondent-driven sampling to recruit MDMA users: a methodological assessment. *Drug and Alcohol Dependence, 78,* 147-57.

Watters, J.K. y Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations, *Social Problems*, *36:* 416-430.