UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Amostragem em Bola de Neve e Respondent-Driven Sampling: uma descrição dos métodos.

Autor: João Osvaldo Dewes
Orientadora: Professora Dra. Luciana Neves Nunes

Porto Alegre, 09 de Dezembro de 2013.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Amostragem em Bola de Neve e Respondent-Driven Sampling: uma descrição dos métodos.

Autor: João Osvaldo Dewes

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professora Dra. Luciana Neves Nunes (orientadora)
Professora Elsa Cristina de Mundstock

Porto Alegre, 09 de Dezembro de 2013.

"A resposta certa, não importa nada:
o essencial é que as perguntas estejam certas."

Mário Quintana

Agradecimentos

Agradeço ao meu pai, João Julio, por ter dado todo o suporte necessário para que eu pudesse me manter e concluir a graduação em estatística. Agradeço à minha mãe, Jucélia, por conseguir aguentar a distância e mesmo assim sempre ter me apoiado nas decisões que eu tomei. Aos meus irmãos Jéssica, Jaqueline, Moisés e Catrina, por terem sempre torcido por mim, mesmo sempre perguntando quando eu iria sair de férias.

Agradeço também aos colegas que ajudaram a tornar o tempo na UFRGS muito mais agradável, dividindo alegrias e decepções. Aos professores que me ajudaram na construção do conhecimento necessário para que eu me tornasse um estatístico.

Por fim, agradeço à professora Luciana por aceitar ser minha orientadora neste trabalho e ter tido tanta paciência comigo, além da professora Elsa que aceitou o convite e se dispôs a avaliá-lo.

Resumo

Há casos em que os pesquisadores se deparam com um certo tipo de população que envolve uma grande dificuldade de se estudar devido à incapacidade de utilizar-se os planos amostrais mais usuais, dada a baixa visibilidade de seus membros, por diversos motivos, sendo alguns deles comportamento ilegal ou socialmente estigmatizado. Essas populações são normalmente denominadas *escondidas* e não possuem um sistema de referências do qual se possa retirar uma amostra probabilística. Apesar dessa dificuldade, em várias delas ocorre um fenômeno que é o de membros dessa população saberem reconhecer outros membros dela, dado que o comportamento que os torna parte da população envolve interações entre seus membros, por motivos tais como, por exemplo, na população de usuário de drogas injetáveis haver o compartilhamento de seringas. A amostragem em bola de neve utiliza-se dessas ligações entre os membros da população para conseguir, partindo de alguns indivíduos membros da população, obter uma amostra dela. O método funciona a partir da indicação por parte de algum indivíduo da população de outros que também fazem parte, e assim sucessivamente, caracterizando-se num formato semelhante ao de uma bola de neve que vai acumulando os flocos de neve ao rolar e se tornando cada vez maior. Como as pessoas costumam se relacionar com outros semelhantes, além de algumas pessoas terem muitas relações e outras serem reclusas sociais, essa seleção costuma trazer viés, primeiro em favor das características das pessoas que começam esse processo e depois das pessoas mais "famosas" em detrimento das que têm menos contatos. Esses fatos tornam impossível saber as probabilidades de seleção de algum indivíduo, caracterizando o processo em um método não-probabilístico. De fato, é de grande interesse dos pesquisadores conseguir fazer inferência para essas populações e dentre as muitas tentativas existe o método *respondent-driven sampling*, que se utiliza de uma ponderação matemática da amostragem em bola de neve para conseguir fazer inferências sobre a população, controlando no modelo estes fatores que influenciam no viés da amostragem em bola de neve. Este trabalho apresenta os dois métodos, primeiro a amostragem em bola de neve, mostrando como funciona a metodologia da sua aplicação e depois o método *respondent-driven sampling*. Mostra-se que o método bola de neve é util para coletar informações de indivíduos de populações escondidas, mas não serve para fazer generalizações, enquanto o método *respondent-driven sampling (RDS)* se mostra muito promissor e consegue medidas interessantes para o pesquisador. Entretanto, o método RDS não possui estimadores confiáveis, sendo seus principais problemas a variância alta destes e as suposições do método que são muito rigorosas e normalmente não atendidas na prática do processo amostral.

Sumário

1) INTRODUÇÃO

Em muitas áreas de pesquisa existe o problema de coleta de informação precisa sobre o comportamento e composição de grupos sociais. Na maioria dos casos, técnicas de amostragem e estimação padrões desenvolvidos ao longo dos últimos 80 anos fornecem meios de obter esta informação. Entretanto, há um número de importantes grupos para os quais estas técnicas não são aplicáveis (Salganik e Heckathorn, 2004).

Estes grupos são definidos como populações escondidas ou difíceis de se encontrar. Podem ser assim definidos por se tratar de membros de uma população-alvo que não sejam distinguíveis da população em geral, ou seja, apesar de estarem lá o pesquisador não sabe quem são, ou por se tratar de uma população em que o comportamento de seus membros envolve um tema sensível que faz com que estes não queiram se revelar. Temas sensíveis envolvem, por exemplo, usuários de drogas injetáveis, imigrantes ilegais ou homens que fazem sexo com homens. São temas sensíveis por envolverem ilegalidade, reprovação social ou ambos os casos. Mas também há casos em que a população-alvo é escondida apesar de seus membros não terem motivos para não se revelarem, como músicos de jazz (Heckathorn e Jeffri, 2003), que não são distinguíveis da população em geral e a maioria não faz parte de associações, além de nem todos costumarem frequentar clubes e festivais de jazz.

Um pesquisador não tem como, por meio de técnicas de amostragem tradicionais, observar uma amostra desses grupos, pois não há como obter um sistema de referências para eles. No método tradicional o pesquisador conhece as probabilidades de seleção de cada elemento da população-alvo devido à existência desse sistema de referências. Ele pode até tentar construir um, mas no caso de uma população escondida ou difícil de encontrar se trata de uma tarefa cara e muitas vezes impraticável. Imagine o exemplo da população de usuários de drogas injetáveis de uma grande cidade, tentar fazer uma lista de todos os usuários é inexequível. Uma alternativa seria buscar fazer amostras em instituições para reabilitação de usuários de drogas, mas também não seria uma amostra aleatória ou representativa da população de usuários de drogas injetáveis, pois provavelmente o comportamento difere entre aqueles que frequentam instituições para reabilitação e aqueles que não frequentam. Temos um exemplo disso num estudo feito em San Francisco (EUA), no qual foi descoberto que usuários de drogas injetáveis fora de programas de tratamento tinham o dobro de probabilidade de estarem infectados com HIV em relação aos que participavam de algum programa de reabilitação (Watters e Cheng, 1987).

Segundo um relatório da Organização Mundial da Saúde (2000), uma das principais fraquezas nos esforços de prevenção do HIV tem sido a dificuldade de monitorar o comportamento e a soro-prevalência de populações em risco, como usuários de drogas injetáveis e homens que fazem sexo com homens, devido a essas influenciarem o modo de propagação do HIV e outras doenças. Então, os pesquisadores têm buscado alternativas para estudar essas populações escondidas, e entre as principais estão a *targeted sampling,* a *time-space sampling* e amostragem em bola de neve (*snowball sampling*).

Na *targeted sampling* os pesquisadores usam diversos métodos de alcance para conseguir coletar uma amostra da população escondida. Normalmente envolve o envio de pessoas para o campo para encontrar e recrutar membro da população. Tem como ponto forte conseguir obter uma amostra com indivíduos de fora da alçada de instituições, mas claramente é um método não-probabilístico no qual não há como se saber a magnitude dos viéses e fazer generalizações para a população. Por exemplo, ao se recrutar usuários de drogas injetáveis, por questões de segurança os pesquisadores só fazem recrutamentos durante o dia. Além disso, usuários que não aparecem em público dificilmente serão encontrados (Salganik e Heckathorn, 2004).

Time-space sampling, também conhecida como *time-location sampling*, é um método que utiliza-se de um trabalho de campo etnográfico para identificar quando os membros da população-alvo se encontram em determinados lugares. Por exemplo, terças, das 2 às 6 da tarde em uma praça específica. A unidades são formadas por esses lugares, então são sorteados alguns e os pesquisadores vão e entrevistam os membros da população que aparecerem por lá. Assim, consegue-se fazer inferências sobre a população nesta situação específica, mas dificilmente o pesquisador saberá qual a diferença entre esse subgrupo e a população-alvo (Salganik e Heckathorn , 2004).

Introduzida incialmente por Coleman (1958) e Goodman (1961), amostragem em bola de neve é um método que não se utiliza de um sistema de referências, mas sim de uma rede de amizades dos membros existentes na amostra. Este tipo de método baseado na indicação de um indivíduo de um ou mais outros indivíduos é também conhecido como método de cadeia de referências. O processo começa de um certo número de sementes, pessoas selecionadas de alguma forma pelo pesquisador e que fazem parte da população-alvo. Essas pessoas, por sua vez, são incumbidas de indicar a partir de seus contatos outros indivíduos para a amostra. Segue-se assim, sucessivamente, até que se alcance o tamanho amostral desejado.

Experiências com o método de amostragem em bola de neve mostraram que ele é efetivo ao penetrar populações escondidas ou difíceis de encontrar. Mas, pela natureza da seleção dos

membros da amostra, que não é aleatória, os pesquisadores não podem confiar neste método para fazer generalizações sobre a população. Uma das causas disso é o fato de que as probabilidades de seleção não são conhecidas. Tem-se conhecimento somente de que a probabilidade de seleção é maior para aqueles com uma rede social maior, enquanto os reclusos sociais tem uma probabilidade pequena de serem selecionados (Salganik e Heckathorn, 2004).

Outra preocupação dos pesquisadores se deve ao fato da escolha das sementes ser muito importante, pois pequenos vieses na escolha destas poderia ser agravado de forma desconhecida conforme seguisse o processo amostral. Estes problemas específicos ocorrem porque o método bola de neve reside fora do mundo da amostragem probabilística tradicional, onde as unidades da amostra são selecionadas com probabilidade de seleção conhecida. Em vez disso, pela falta de sistema de referências e pelas probabilidades de seleção desconhecidas, amostras em bola de neve são consideradas não-probabilísticas ou de conveniência, "que só podem ser avaliadas subjetivamente" (Kalton, 1983). Berg (1988) resumiu bem as críticas a este método quando escreveu:

> "De regra, uma amostra em bola de neve será fortemente viesada em favor da inclusão daqueles que tem muitas interrelações ou são pareados com um grande número de indivíduos. Na ausência de conhecimento das probabilidades individuais de inclusão nas diferentes ondas da amostragem em bola de neve, estmação não-viesada não é possível."

Assim, sociólogos, pesquisadores de saúde pública e estatísticos chegaram à conclusão de que a amostragem em bola de neve seria sim promissora, especialmente no estudo de populações escondidas ou difíceis de encontrar, mas poderia ser tão viesada que não deveria ser usada para fazer estimativas confiáveis.

Desenvolvido por Heckathorn (1997), para estudos de prevenção do HIV, o método *Repondent-Driven Sampling* (RDS) se baseia no método amostragem em bola de neve e utiliza-se de um modelo matemático que pondera os indivíduos da amostra conforme seu grau de relações sociais, tentando eliminar o viés de seleção e obter estimativas confiáveis nos estudos de populações escondidas ou difíceis de encontrar.

No *Web of Science* foram encontrados 935 artigos com as palavras-chave *snowball sampling*, sendo que 506 destes são dos últimos 5 anos. Para artigos brasileiros, estes números são de 42 e 25, respectivamente. Quando as palavras-chave foram *respondent-driven sampling*, os resultados foram 480 artigos, sendo 389 nos últimos 5 anos, e no Brasil sendo 19 e 16, respectivamente. A

popularidade destes métodos se dá pela rapidez de aplicação e custo-efetividade ao aplicá-los para amostragem de populações escondidas ou difíceis de encontrar.

Dado esse crescente e a larga utilização dos métodos de amostragem em bola de neve e *respondent-driven sampling*, o objetivo deste trabalho é apresentar e descrever as técnicas, além de definir como fazer uso adequado delas e obter resultados que sejam condizentes com a realidade.

2) AMOSTRAGEM EM BOLA DE NEVE

Amostragem em bola de neve é um método tipicamente utilizado com populações raras ou desconhecidas. Membros destas populações não foram todos identificados previamente e são mais difíceis de encontrar ou contatar do que populações conhecidas (Coleman, 1958; Goodman, 1961; Spreen, 1992).

2.1) Motivação para o uso desta técnica

Populações raras ou desconhecidas não possuem uma lista de seus membros e construir uma é tarefa difícil ou impraticável, incluindo um alto custo financeiro e de tempo para aplicação (Salganik e Heckathorn, 2004).

2.2) Quando pode ser usada

O método de amostragem em bola de neve pressupõe que há uma ligação entre os membros da população dado pela característica de interesse, isto é, os membros da população são capazes de identificar outros membros da mesma. Por exemplo, moradores de rua provavelmente conhecem outros moradores de rua e podem levar o pesquisador a encontrá-los. Já em outro caso, sonegadores de impostos não têm qualquer ligação aparente entre si, o que torna improdutivo o uso do método de amostragem em bola de neve (Faugier e Sargeant, 1997).

2.3) Método de aplicação

O primeiro passo no método de amostragem em bola de neve é encontrar indivíduos pertencentes à população-alvo do estudo. Esses indivíduos vão ser a *semente* da amostra, aqueles que darão origem a todos os indivíduos amostrados. Este é um passo muito importante, pois se essa *semente* não for bem selecionada a amostra não conseguirá atingir toda a variabilidade da população. Na maioria dos casos, as sementes costumam ser as pessoas mais acessíveis aos pesquisadores, mas é recomendável que se faça um estudo maior sobre onde podem ser encontrados indivíduos da população, para encontrar indivíduos que produzam uma amostra menos viesada

(Snijders, 1992). Partindo-se desde associações profissionais, de organizações de reabilitação, pontos de encontro e assim por diante, o importante é que se busque o máximo de referências.

A partir da *semente* começa o processo da bola de neve. Esse primeiros indivíduos são considerados a *onda zero.*

- Inicia-se o processo pedindo a cada semente que indique o contato de *n* outros indivíduos que eles consideram ser membros da população-alvo.
- A *onda um* é formada pelos contatos indicados pelos indivíduos da *onda zero* que fazem parte da população-alvo e que não fazem parte da *onda zero*.
- A *onda dois* é formada pelos contatos indicados pelos indivíduos da onda um que fazem parte da população-alvo e que não fazem parte da *onda zero* nem da da *onda um*.
- O processo segue até que o tamanho de amostra desejado seja alcançado ou então quando uma nova onda não produza um determinado número de contatos novos.

2.3.1) Observações

A quantidade indivíduos envolvidos na *onda zero* depende normalmente da capacidade do pesquisador de conseguir reunir membros da população-alvo dispostos a começar o processo e indicar novos contatos para o estudo. Como são populações difíceis de encontrar normalmente a parte mais difícil do estudo acontece aqui. Frequentemente estudos requerem algum "conhecimento de pessoas de dentro" prévio para conseguir recrutar os primeiros respondentes (Atkinson e Flint, 2001).

Se a quantidade de contatos indicados por cada indivíduo for pequena, digamos 1 ou 2, a cadeia de contatos costuma morrer em poucas ondas. Por outro lado, se for muito grande, costuma favorecer aqueles com mais contatos. Limitar esse número para ficar entre 3 e 6 contatos tem se mostrado na prática ser eficiente em resolver ambos os problemas.

Normalmente se termina o processo amostral ao chegar num tamanho de amostra definido antes da pesquisa como alvo, ou então quando se atinge uma estabilidade, ou seja, quando poucos novos contatos são acrescentados. Considerar o que são poucos contatos fica a cargo do pesquisador, mas pode ser uma proporção, por exemplo, menos que 5% dos contatos indicados na onda que tenham respondido.

2.4) Estimação e generalizações

Desde o surgimento do método de amostragem em bola de neve busca-se adaptá-lo de modo a obter estimativas generalizáveis para a população. Goodman (1961) definiu toda a formulação matemática do método. Apesar disso, nenhum aspecto prático de como deveria ser aplicado o método de modo a cumprir as suposições deste foi desenvolvido. Não havia a consideração da parte que envolvia as pessoas, nem as ligações entre elas, o caráter social do método. Biernacki e Waldorf (1981) foram os primeiros a dar atenção ao método e colocá-lo como foco de um estudo, sendo que antes amostragem em bola de neve era meramente citada rapidamente pelos pesquisadores na metodologia como se os seus problemas fossem evidentes e que não precisassem de maior aprofundamento. Atkinson e Flint (2001) citaram a impossibilidade ainda de se obter estimativas confiáveis, e portanto recomendam o lado qualitativo do método.

Os estudos qualitativos servem para verificar o comportamento da população em estudo e como ele é afetado pela estrutura da rede social na qual ela está inserida. A sugestão é de que deva-se focar no lado metodológico, controlando o processo de amostragem para conseguir extrair o máximo de informação sobre as características da população em estudo. Partindo do ponto de que este trabalho seja bem feito, pode-se então utilizá-lo na comparação com outros estudos semelhantes (Bienarcki e Waldorf, 1981).

Deve-se ter cuidado das limitações dos resultados encontrados em estudos que utilizam a amostragem em bola de neve para não se chegar a conclusões incorretas. Ao apresentar os resultados deve estar bem claro em quais condições estes foram obtidos, para que outros pesquisadores, quando lerem, obtenham respostas que são verdadeiras e tenham consciência de para que e quem elas são válidas.

Se o objetivo do estudo é exploratório, qualitativo e descritivo, a amostragem em bola de neve oferece claramente vantagem em obter informações de populações escondidas ou difíceis de encontrar que seriam de outro modo muito difíceis de se coletar (Hendricks e Blanken, 1992).

2.5) Pontos fortes

O método de amostragem em bola de neve permite ao pesquisador encontrar populações que ele não conseguiria através de outros métodos. Principalmente no caso de populações que são caracterizadas por comportamentos ou histórico que as fazem não querer aparecer ou ter sua condição revelada. Pelo método de recrutamento ser feito através da indicação de outras pessoas

que também são membros da população o processo é facilitado, pois normalmente envolve uma relação de confiança que não existiria com um pesquisador desconhecido fazendo esta abordagem (Biernacki e Waldorf, 1981).

Além disso, já é conhecido por ser um processo que é barato e custo-eficiente quando comparado a métodos alternativos de recrutamento de populações escondidas ou difíceis de encontrar como *targeted sampling* ou *time-space sampling.* O processo de amostragem em bola de neve precisa de menos planejamento e pessoas, quando comparado a essas técnicas (Salganik e Heckathorn, 2004).

Devido ao fato de ser eficiente ao penetrar populações escondidas ou difíceis de encontrar, como por exemplo pessoas que apresentam comportamento de risco em relação ao HIV, como usuários de drogas injetáveis ou homens que fazem sexo com homens, apresenta o fator positivo de ser utilizado para disseminação de informações de cuidados e comportamentos adequados para auxiliar no bem-estar e saúde dessas populações, além de diminuir a propagação do HIV e outras doenças.

2.6) Problemas

A primeira complicação do método começa justamente na parte em que ele é útil, ou seja, no que diz respeito à coleta das informações de populações difíceis de encontrar. Partindo do fato de que para começar a cadeia de referências que vai gerar a amostra é necessário obter os indivíduos que serão a base desse processo, ou as *sementes*, a visibilidade social da população-alvo é um dos primeiros problemas do pesquisador ao pensar em começar uma amostragem em bola de neve (Heslin, 1972). Policiais, enfermeiros e professores têm alta visibilidade social, e embora possa vir a ser difícil acessar suas populações, não há dificuldade de saber onde encontrá-los. Outras possíveis populações de estudo, por outro lado, por motivos sensíveis em relação ao comportamento moral, social ou legal tem pouca visibilidade e podem ocasionar sérios problemas para se encontrar potenciais respondentes (Biernacki e Waldorf, 1981).

Como já mencionado anteriormente, por ser um método não-probabilístico, não se pode fazer nenhuma inferência sobre a população, o que limita as conclusões que os pesquisadores possam fazer. Também há pouco controle do pesquisador sobre o método amostral, os novos indivíduos da amostra dependem basicamente dos que já estão presentes na amostra. Também há a questão dos indivíduos que são selecionados terem uma homogeneidade com aqueles que os indicaram, compartilhando traços e características, podendo até levar à amostra a ser apenas um subgrupo da

população-alvo do estudo (Faugier e Sargeant, 1997).

Normalmente a amostra inicial, a *semente,* é selecionada por uma amostra de conveniência, o que já pode trazer viés. Outras fontes de viés conhecidas são o voluntariado e o mascaramento.
O voluntariado se refere à vontade de participar e colaborar com o estudo, e isto varia de um indivíduo para outro, trazendo viés na seleção destes. O mascaramento se caracteriza em uma proteção do indivíduo quanto a seus amigos e parentes, não querendo revelar que estes fazem parte da população, normalmente quando envolve um comportamento estigamatizado, como uso de drogas. Embora seja um problema maior em população escondidas ou difíceis de encontrar, o mascaramento também pode ocorrer em qualquer outro tipo de estudo ou população (Salganik e Heckathorn, 2004).

Membros da população com uma maior rede de contatos e maior visibilidade social têm maior chance de serem indicados (Biernacki e Waldorf, 1981; Henslin, 1972). Reciprocamente, aqueles com menores redes de contatos, ou indivíduos isolados, podem ser omitidos da amostra por terem menor chance de serem mencionados por algum outro membro da população (Van Meter, 1990).

3) RESPONDENT-DRIVEN SAMPLING

Amostragem dirigida pelo participante ou, do inglês, *Respondent-Driven Sampling* (RDS) (Brignol, 2013:56) é um método que combina amostragem em bola de neve com um modelo matemático que pondera a amostra para compensar o fato de ter sido coletada de uma maneira não-aleatória. Este modelo é baseado em uma síntese e extensão de duas área da matemática: teoria de cadeias de Markov e teoria das redes viesadas.

3.1) Propósito

Tem o mesmo objetivo da amostragem em bola de neve, dado que é baseado neste método de seleção. Alcançar populações, que por motivos diversos são escondidas ou difíceis de encontrar, seja por estigmatização social, ilegalidade ou outras razões quaisquer.

3.2) Razão de utilização

A amostragem em bola de neve já é amplamente utilizada em diversos meios, mas seu método de aplicação tem uma grande variação dependendo de quem o aplica, quais as condições do estudo e quais os objetivos, com a intenção de facilitar a aplicação por parte dos pesquisadores. Alguns exemplos de adaptações e variações da estratégia de recrutamento da amostragem em bola de neve podem ser encontrados em Sadler et al. (2010).

O método *respondent-driven sampling* (RDS), criado em 1997 por Heckathorn, traz uma sistematização intrínseca na sua aplicação, tendo regras e suposições bem definidas, desde a população à qual pode ser aplicada até o processo de análise dos dados, embora ainda forneça bastante espaço para a criatividade do pesquisador para resolver problemas específicos do seu estudo.

3.3) Diferenças em relação ao método bola de neve

Incentivo: o método de amostragem em bola de neve normalmente oferece uma recompensa para participação. No método RDS há um dupla recompensa, tanto para participação quanto para cada recrutamento.

Recrutamento: enquanto no método de amostragem em bola de neve o pesquisador solicita ao respondente que liste outros membros da população-alvo, no método RDS o próprio indivíduo é responsável por recrutar outros.

3.4) Limitações

Como no caso de amostragem em bola de neve, o método RDS só é adequado quando a população-alvo tem por critério de inclusão alguma característica que gere algum tipo de ligação entre os membros dela como, por exemplo, usuários de drogas quando compartilham ou compram drogas, ou quando atividades sexuais de alto risco acontecem.

Portanto, este método não é adequado para fazer amostras de grande escala territorial. O tamanho da área na qual o processo amostral pode ser considerado efetivo depende do padrão de extensão geográfica dos contatos, que por sua vez depende da disponibilidade de transporte para os respondentes.

3.5) Suposições

1) Os respondentes devem ter relacionamentos recíprocos com os indivíduos que eles sabem serem membros da população.
2) Cada respondente pode ser alcançado por qualquer outro respondente através de uma série de nós da rede de contatos, isto é, a rede de contatos forma um único componente.
3) A amostragem é com reposição.
4) Os respondentes podem reportar com precisão o tamanho da sua rede de contatos pessoal ou equivalentemente, seu grau.
5) O recrutamento de seus "iguais" é uma seleção aleatória dos "iguais" do recrutador.

As três primeiras suposições especificam as condições necessárias para que a RDS seja um método de amostragem apropriado para a população. Primeiro, de modo que o recrutamento possa ocorrer, o respondente deve ter acesso aos outros membros da população e ser capaz de identificar quais deles se qualificam para o recrutamento. Em adição, estimativas de RDS são baseadas em uma estrutura de rede na qual os nós são recíprocos (Heckathorn, 2002). Formalmente, a reciprocidade significa que se A recruta B, então deve haver uma probabilidade diferente de zero

que B pudesse recrutar A. Consequentemente, o delineamento de pesquisa da RDS inclui meios de encorajar os indivíduos a recrutar seus conhecidos ou amigos em vez de estranhos, através de recompensa por sucessos no recrutamento e, além disso, oferecendo apenas um número limitado de chances de recrutamento, tornando-os valiosos (Heckathorn, 1997). Isto é, quando os respondentes são limitados a um certo número de recrutamentos que podem fazer para receber a recompensa, tendem a pensar duas vezes antes de desperdiçá-los com estranhos. Segundo, assume-se que a população forme um único componente (Salganik e Heckathorn, 2004). Em outras palavras, todos da população-alvo devem ser alcançáveis por um único respondente através de um conjunto finito de nós da rede de contatos. Em uma rede aleatória, um único componente se forma quando os graus individuais são grandes quando comparados ao logaritmo natural do tamanho da população (Bollabás, 1985; Watts e Strogatz, 1998). Quando é permitido aos respondentes recrutar não somente aqueles com os quais tem um relacionamento especial, mas também amigos e conhecidos que eles sabem fazer parte da população-alvo, então os graus individuais são maiores que aqueles geralmente necessários para que a rede forme um único componente (Heckathorn, 2007). Adicionalmente, já que redes sociais nunca são realmente aleatórias, um requerimento mínimo é de que não haja alguma barreira estrutural ou social que segregue completamente um subgrupo da população. Por exemplo, RDS não pode ser usada em uma amostra através de castas em uma cultura onde interação entre castas é proibida. Terceiro, a teoria estatística para estimação na RDS é baseada num esquema de amostragem com reposição (Salganik e Heckathorn, 2004). Consequentemente, a fração amostral deve permanecer pequena o suficiente para que o modelo com reposição seja considerado apropriado (Heckathorn, 2007).

As duas últimas suposições são potencialmente as mais problemáticas. A suposição quatro requer que os respondentes tenham a capacidade de fornecer informações precisas quanto ao tamanho de sua rede social, uma tarefa que costuma ser difícil até para especialistas em rede sociais. A suposição cinco assume que o padrão de recrutamento reflete a composição da rede de contatos pessoal dentro da população-alvo. Isto é dizer que os respondentes recrutam aleatoriamente de suas redes pessoais (Heckathorn, 2002).

3.6) Método

1) São recrutados alguns indivíduos que servem de *sementes.*

2) Estes indivíduos são entrevistados. Ao completarem a entrevista, lhes é oferecido incentivo financeiro para recrutar outros indivíduos da mesma população. Especificamente, eles recebem cupons de recrutamento e lhes é dito que para cada pessoa recrutada que apareça e complete uma entrevista será dada ao recrutador uma recompensa em dinheiro ou de alguma outra forma.

3) Todos os indivíduos que são recrutados recebem a mesma oportunidade de recrutar, como aconteceu no caso daqueles que eram as sementes. Eles recebem uma recompensa pela participação no estudo e outra para cada recrutamento realizado com sucesso. Oferecendo-se incentivos adequados, este mecanismo cria um sistema que vai expandindo a cadeia de referências, na qual os indivíduos recrutam outros mais, que vão recrutar outros, e daí por diante, de onda para onda. Para assegurar que uma ampla gama de indivíduos tenha oportunidade de recrutar, prevenir o surgimento de recrutadores semi-profissionais e batalhas sobre direito de recrutamento, cada respondente deve ser limitado a um número de cupons iniciais.

4) As características que definem que o sujeito é membro realmente da população precisam ser objetivamente verificáveis, para que os respondentes não recrutem indivíduos que não são membros da população-alvo a fim de receber o incentivo financeiro. Como exemplo, temos o caso de usuários de drogas injetáveis, onde é possível se verificar objetivamente a utilização ou não por parte do respondente.

5) Existem os casos de duplicação de sujeito, quando um respondente tenta participar do estudo sob múltiplas identidades, e casos de personificação, quando um respondente tenta se passar por outro, talvez como meio de ganhar a recompensa de recrutamento deste. Estes potenciais problemas podem ser superados usando uma base de dados de identificação de indivíduos, gravando características físicas como sexo, gênero, idade, etnia, altura, cicatrizes, tatuagens, e algumas outras medidas biométricas.

6) Visar subgrupos específicos dentro da população-alvo pode ocorrer através de incentivos de direcionamento, isto é, bônus para recrutamento de categorias ou indivíduos específicos.

7) A amostragem pode ser encerrada quando a comunidade alvo está saturada ou então quando um tamanho alvo mínimo de amostra tenha sido alcançado e a composição da amostra tenha se tornado estável em relação às variáveis que a pesquisa foca.

3.7) Escolha das sementes

Existem recomendações quanto à seleção de *sementes:* primeiro, sementes devem ser diversificadas com respeito aos fatores que mais fortemente determinam a formação de ligações sociais dentro da população. Tipicamente, estas são características demográficas básicas tais como raça, etnia, religião, casta, status social e idade. Segundo, porque muitas ligações sociais são formadas baseadas em proximidade tais como viver na mesma rua ou trabalhar no mesmo local, as *sementes* devem ser retiradas de uma variedade de áreas geográficas ocupadas pela população-alvo. Finalmente, *sementes* devem ser estrelas sociométricas, isto é, indivíduos que mantém muitas ligações sociais e são altamente considerados dentro da população-alvo. Além disso, devem ser comprometidos com os objetivos do estudo. Tais indivíduos podem com mais facilidade promover a participação de outras pessoas e acelerar o recrutamento. Paradoxalmente, cuidadosa seleção de *sementes* acelera o crescimento das cadeias de recrutamento e, portanto, aceleram o ponto no qual a seleção das sementes se torna irrelevante, por consequência ajudando a reduzir o viés (Ramirez-Valles, 2005).

3.8) Dados necessários para a análise RDS

Embora os dados necessários para análise RDS sejam mínimos, há três peças de informação que são essenciais para a análise (ela NÃO PODE SER FEITA sem estes campos para cada respondente).

- Tamanho da rede de contatos pessoal (grau do respondente) – número de pessoas que o respondente conhece dentro da população-alvo;
- Número de série do respondente – número de série do *cupom* com o qual o indivíduo foi recrutado;
- Números de série recrutados pelo respondente – números de série dos *cupons* recebidos pelos indivíduos para recrutar outras pessoas.

3.9) Cadeias de Markov

Uma cadeia de Markov é formada por realizações de um processo estocástico, onde a probabilidade de um evento ocorrer depende dos eventos anteriores. Os possíveis eventos deste processo são chamados de estados do processo e todos os estados formam o espaço de estados. A cada realização do processo existe a probabilidade de um estado passar para outro e estas probabilidades são chamadas de probabilidades de transição. A matriz que contém as probabilidades de um estado passar para outro é chamada de matriz de transição. Quando vista analiticamente, uma aplicação de RDS cria um processo estocástico no qual cada característica social do recrutador afeta as características dos recrutados. Por exemplo,
podemos considerar que a probabilidade de um indivíduo de um tipo, por exemplo, etnia ou sexo, recrutar outro, depende apenas de sua etnia ou sexo e não depende de quem o recrutou anteriormente, assim formando um processo sem memória, que se caracteriza como uma cadeia de Markov de primeira ordem, onde a probabilidade de se sair de um estado e ir para outro depende apenas do estado atual. Consideremos também que partindo-se de qualquer tipo pode-se chegar a alguém de qualquer outro tipo, de um número finito de tipos, com probabilidade positiva. Assim temos uma cadeia ergódica, e considerando que existe uma probabilidade positiva do indivíduo recrutar dentro do próprio tipo, ou seja, permanecer no mesmo estado, temos uma cadeia não-cíclica, e portanto, regular.

Uma propriedade importante das cadeias de Markov regulares é que depois de passado algum tempo do processo, as probabilidades de se estar em algum determinado estado não depende do estado inicial, ou seja, ela converge para uma distribuição-limite, que são as probabilidades de se estar em algum estado depois de um tempo que tende ao infinito: lei dos grandes números para cadeias de Markov (Kemeny e Snell, 1960:73). Considerando no caso de RDS, podemos supor então que, se esse recrutamento se comporta como uma cadeia de Markov regular de primeira ordem, independente do tipo inicial do recrutador, ou semente, chegará a um ponto em que a probabilidade de ser recrutado algum tipo independe do tipo inicial, ou seja, chegamos na probabilidade verdadeira de alguém ser recrutado, que no caso, seria a proporção a qual esse tipo de indivíduo ocorre na população. Algo que assim ajudaria a eliminar os problemas de viés da seleção não-aleatória de *sementes* feita no começo do processo de amostragem.

Também pela lei dos grandes números para cadeias de Markov, tem-se o que Kemeny e Snell (1960:72) descreveram como “um tipo de convergência muito rápida”. Esta conclusão foi baseada

em dedução constatando que a convergência para a distribuição-limite ocorre numa taxa de progressão geométrica. Mas levando em consideração que a amostra geral é constituída tanto pela onda atual quanto as precedentes, este processo na verdade é mais lento, embora ainda rápido. Também varia conforme o número de indivíduos recrutados por cada respondente, um número maior representa uma convergência mais rápida para um recrutamento na proporção verdadeira da população.

Essas conclusões geraram dois teoremas por parte de Heckathorn (1997):

1) Conforme o processo de recrutamento continua, de onda para onda, uma mistura equilibrada de recrutados que é independente das características do sujeito ou conjunto de sujeitos dos quais o recrutamento começou eventualmente será obtida.
2) O conjunto de indivíduos gerados por um processo *respondent-driven sampling* alcança o equilíbrio à uma taxa rápida (isto é, geométrica).

Se o recrutamento tomar a forma do processo de Markov mais simples, isto é, uma cadeia linear começando com uma única *semente,* a implicação dos teoremas (1) e (2) é de que as cadeias de recrutamento devem ser longas. Mas isso implica no problema, de que dado a incapacidade de um indivíduo recrutar novos membros, as cadeias costumam morrer em menos de 3 passos (Klovdahl, 1989). Portanto, a abordagem usada na maioria dos métodos de cadeia de referência permite aos respondentes indicarem múltiplas referências, formando assim um recrutamento ramificado. Deste modo, uma falha individual de recrutar não "mata" a cadeia e dado o devido tempo e recursos, cadeias de qualquer tamanho são virtualmente realizáveis. Outra vantagem dos múltiplos contatos é a de diminuir o viés da escolha inicial de indivíduos porque os recrutados se tornam mais distantes socialmente das *sementes*.

Um problema desse processo de recrutamento ramificado é de que ele não corresponde à estrutura linear suposta pelo modelo de cadeias de Markov. Uma possível solução seria analisar cada uma das sementes como tendo seu próprio processo linear e o processo final se tratar de uma combinação linear. Também para avaliar se o modelo de cadeias de Markov se ajusta aos dados pode ser feita uma comparação empírica, comparando a composição da amostra com a composição teórica que seria obtida se o processo amostral correspondesse a um processo de Markov, isto é, o equilíbrio. Por exemplo, em uma aplicação prévia de RDS (Heckathorn, 1997:188-189), uma grande discrepância (17%) foi encontrada entre o equilíbrio computado teoricamente e a média

amostral, e examinando-se foi descoberto que o recrutamento não era ergódico. O problema foi resolvido dividindo-se a amostra em duas sub-amostras, cada qual por sua vez sendo ergódica.

3.10) Estimação

Existem dois estimadores principais de *respondent-driven sampling,* tendo o primeiro, RDS I sido desenvolvido por Heckathorn (1997) e o segundo, RDS II, desenvolvido por Volz e Heckathorn (2008), Esses estimadores estimam a proporção dos grupos dentro da população.

3.10.1) RDS I

O estimador RDS original usa um processo de dois estágios onde os dados são usados para fazer inferência sobre a estrutura da rede e então essas inferências são usadas para fazer inferências sobre a população. Especificamente, foi mostrado que sob certas suposições as probabilidades de transição entre os grupos, a probabilidade de sair de um tipo para algum outro, por exemplo, um respondente homem recrutar uma respondente mulher, estimadas pelas probabilidades de transição da amostra, podem ser usadas junto com o grau médio do grupo, o grau sendo o número de contatos da um dado indivíduo na população-alvo, para calcular estimativas não-viesadas da proporção populacional de dados baseados em uma rede de contatos (Salganik e Heckathorn, 2004). Sob a suposição de reciprocidade, o número de ligações ou recrutamentos do grupo X para o grupo Y se iguala ao número de ligações ou recrutamentos do grupo Y para o grupo X. Entretanto, em uma amostra finita, e este costuma ser o caso, é usual haver um número de recrutamentos de um grupo A para um outro B, e esse número não ser igual no caminho contrário, B para A.

Assim, Heckathorn (2002) melhorou a estimativa das ligações inter-grupos através de um processo conhecido como *data-smoothing* , no qual cada número de recrutamentos inter-grupos é ponderado de modo tal que a matriz de nível do grupo de quem recrutou quem, denominada *matriz de recrutamento* , é simétrica. As probabilidades de transição baseadas na matriz de recrutamento com *data-smoothing* são então combinadas com a estimativa do grau para calcular uma estimativa do tamanho proporcional do grupo, $\widehat{P_x^{RDSI}}$.

$\widehat{P_X^{RDSI}} = \frac{\widehat{S_{YX}}\widehat{D_Y}}{\widehat{S_{YX}}\widehat{D_Y} + \widehat{S_{XY}}\widehat{D_X}}$, onde

$\widehat{S_{XY}}$ é a proporção com *data-smoothing* de recrutamentos do grupo X para o grupo Y,

$\widehat{S_{YX}}$ é a proporção com *data-smoothing* de recrutamentos do grupo Y para o grupo X,

$\widehat{D_X}$ é a estimativa do grau médio do grupo X e

$\widehat{D_Y}$ é a estimativa do grau médio do grupo Y.

Intervalos de confiança para RDS I são estimados usando um algoritmo de *bootstrap* (Heckathorn, 2002; Salganik, 2006). O algoritmo gera uma reamostragem das observações baseadas na matriz de transição amostral. Isto é, se 70% dos recrutamentos do tipo A são outros A's e a observação atual é do tipo A, o algoritmo vai gerar um A como próxima observação na reamostragem com probabilidade 0,7. Este processo continua até que a reamostragem alcance o tamanho original da amostra. As estimativas RDS I são então calculadas e o processo é repetido até que um número especificado de reamostragens tenha sido alcançado. As caudas do intervalo de confiança são tomadas da distribuição dessas estimativas *bootstrap*. Isto é, o limite superior de um intervalo de 95% de confiança é definido como sendo o ponto acima do qual recaem 2,5% das estimativas *bootstrap,* analogamente para o limite inferior. Consequentemente, o algoritmo *bootstrap* permite intervalos de confiança não-simétricos, e além disso, não fornece uma estimativa direta da variância.

3.10.2) RDS II

Usando a abordagem de estimativa baseada na probabilidade, Volz e Hecaktorn (2008) inferem que uma amostra baseada numa rede de contatos selecionará os indivíduos na população com probabilidade proporcional ao grau e derivam um novo estimador RDS, $\widehat{P_X^{RDS\,II}}$:

$$\widehat{P_X^{RDS\,II}} = \left(\frac{n_X}{n.}\right)\left(\frac{\widehat{D.}}{\widehat{D_X}}\right)$$, onde

n_X é o número de respondentes no grupo X,

$n.$ é o número total de respondentes,

$\widehat{D_X}$ é o grau médio do grupo X, e

$\hat{D}.$ é o grau médio geral.

Essencialmente, a estimativa de $\widehat{P_X^{RDS\,II}}$ é a proporção da amostra, $\frac{n_X}{n}$, ponderada pela correção dos efeitos da rede de contatos, $\frac{\widehat{D.}}{\widehat{D_X}}$. Uma vantagem do estimador RDS II é que ele é calculado diretamente dos dados, removendo o passo do intermediário de fazer inferência sobre estrutura da rede, necessária no RDS I. O RDS II também permite a análise de variáveis contínuas, enquanto o RDS I não permite.

Volz e Heckathorn (2008) mostram que as estimativas RDS I e RDS II convergem quando a matriz de recrutamento é simétrica. Assim, quando o *data-smoothing* é utilizado, um procedimento que é recomendado para todas as análises RDS e que é o padrão no programa RDSAT (Volz et al., 2012), os estimadores RDS I e RDS II produzem resultados equivalentes. A grande diferença é que a abordagem matemática usada para calcular as estimativas no RDS II permite o cálculo analítico da variância, enquanto a abordagem do RDS I não permite tal solução.

Os limites do intervalo de confiança para estimativas RDS II são baseados no estimador da variância do RDS II (Volz e Heckathorn, 2008):

$$Var\left(\widehat{P_X^{RDS\,II}}\right) = \widehat{V}_1 + \frac{\widehat{P_X^{RDS\,II}}^2}{n.}\left(\left(1-n.\right) + \frac{2}{n_X}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\widehat{S}.^{i-j}\right)_{XX}\right)$$

onde

$$\widehat{V}_1 = \frac{\widehat{Var(Z_i)}}{n.} = \frac{1}{n.(n.-1)} \sum_{i=1}^{n} \left(Z_i - \widehat{P_X^{RDS\,II}} \right)^2$$

e

$$Z_i = d_i^{-1} \widehat{D}. I_X(i)$$

onde d_i é o grau do respondente i, $\widehat{S}.$ é a matriz de probabilidades de transição, e $I_X(i)$ é a função indicadora que o toma o valor 1 se $i \in X$ e 0, caso contrário. Embora a estimativa seja viesada, Volz e Heckathorn (2008) descobriram que se aproxima das estimativas não-viesadas de suas simulações.

3.10.3) Homofilia

É uma medida estimada também importante, pois trata do quanto um grupo recruta dentro do próprio grupo. Varia de -1 a 1, sendo que quando é 1 os indivíduos daquele grupo recrutam apenas dentro do próprio grupo, quando é 0 significa que os indivíduos recrutam qualquer um independente do grupo, e quando é -1 significa que os indivíduos só recrutam fora do seu grupo. Homofilia se trata do comportamento das pessoas de se relacionarem com outros que são parecidos com elas, por exemplo, tem a mesma idade, pertencem à mesma classe social, mesma etnia, mesmas preferências e assim por diante. Heterofilia representa o oposto, quando os indivíduos se relacionam com pessoas que diferem de si em algum critério, sendo o exemplo mais óbvio, relacionamentos heterossexuais. Homofilia maior que 0,3 é definida como intermediária enquanto valor menor que -0,3 é classificada como forte heterofilia. Um exemplo do cálculo da homofilia: temos dois grupos, A e B. Se A recruta B 1/3 das vezes e A 2/3 das vezes, a homofilia é igual a 2/3 menos 1/3, ou seja, 1/3. Isso significa que A recruta 1/3 das vezes dentro do próprio grupo enquanto no resto do tempo recruta aleatoriamente entre A e B (Heckathorn, 2002).

3.11) Softwares Disponíveis Para RDS

Como principal ferramenta para a análise RDS se apresenta o software RDSAT (Respondent Driven Sampling Analysis Tool). Ele está disponível gratuitamente para baixar no site *respondentdrivensampling.org* e no mesmo site há também um manual que explica como utilizá-lo. Esse software pode importar bancos de dados do SAS, SPSS e Stata, desde que se tomando os devidos cuidados com a sua formatação especial que precisa da identificação dos recrutamentos. Para o controle dos cupons e dos incentivos dados aos respondentes existe também o software RDSCM (Respondent Driven Sampling Cupoun Manager), também disponível para baixar. Para construção dos gráficos da cadeias de recrutamento, existem os softwares NetDraw e UCINET. No R também já existe um pacote chamado RDS para fazer as análises de amostras provenientes de *respondent-driven sampling.*

3.12) Exemplo de Análise RDS

Para ilustrar a aplicação de uma análise aos dados de uma amostra feita atráves do método RDS será utilizado o banco de dados que vem como exemplo no programa RDSAT, versão 7.1.38. O banco chamado *nyjazz.rds* é derivado de uma amostragem feita através de RDS de músicos de jazz de Nova York (Heckathorn e Jeffri, 2001). As variáveis contidas no banco são idade, sexo, raça, se faz parte do sindicato, se já teve seu trabalho tocado no rádio e o grau declarado, ou seja, quantos contatos considera conhecer na população-alvo. Foi um estudo feito para testar RDS num contexto de população-alvo não estigmatizada.

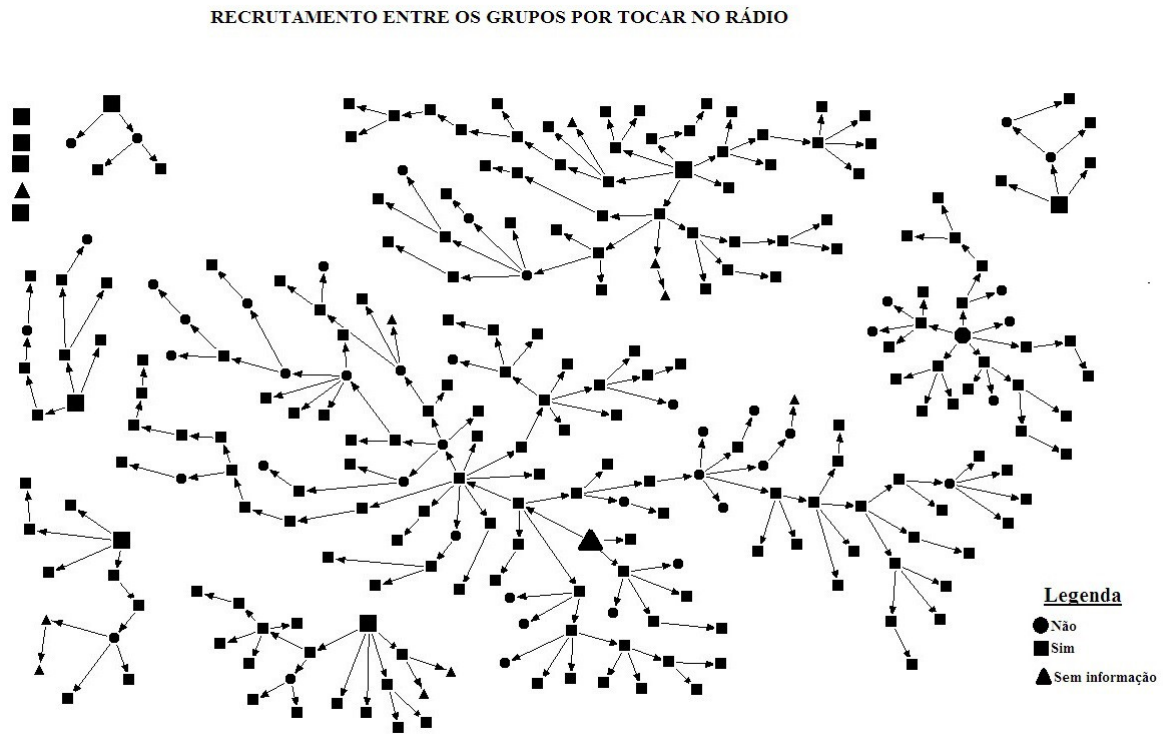


Figura 1: Recrutamento por tocar no rádio
Observação: as figuras geométricas maiores representam as sementes

Neste estudo haviam 13 sementes e podemos ver pelo grafo da figura 1 que cinco destes não conseguiram recrutar algum respondente. A proporção de músicos que já tiveram seu trabalho tocado no rádio na amostra é bem maior do que os que não tiveram. Inclusive, apenas um músico que fazia parte da semente não tinha tocado no rádio. Apesar disso, da cadeia de recrutamentos gerada por ele, temos que a grande maioria é formada por aqueles que já tocaram. Vemos também que alguns não recrutaram, e outros geraram cadeias pequenas, enquanto uma semente foi responsável por aproximadamente metade dos respondentes obtidos na amostra, mesmo ele recrutando apenas três dos seus sete recrutamentos possíveis neste estudo, também não se sabe qual a classificação desta semente.

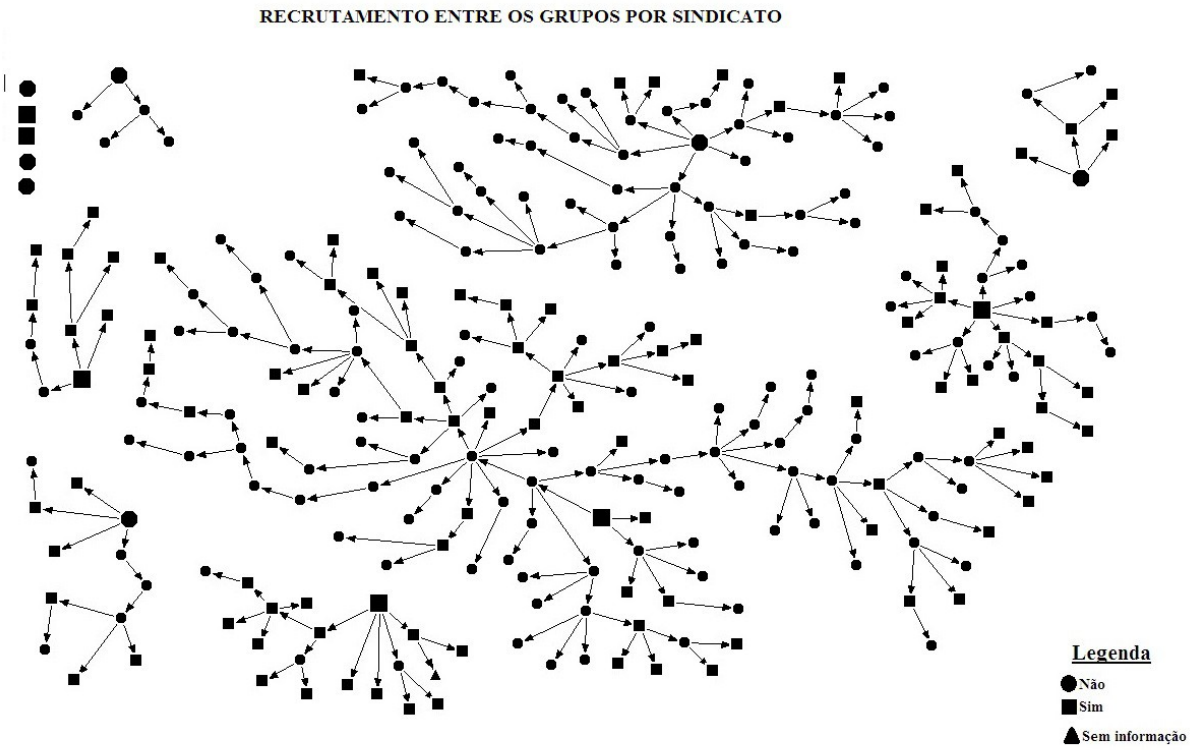


Figura 2: Recrutamento por fazer parte do sindicato
Observação: as figuras geométricas maiores representam as sementes

Em relação à participação no sindicato temos uma divisão mais equilibrada entre os grupos, o que pode se perceber inclusive na *semente* que contava com seis participantes do sindicato e sete que não faziam parte do sindicato. Aqui descobrimos que o responsável pela maior cadeia de contatos faz parte do sindicato, enquanto o da segunda maior cadeia não. Ainda pode-se perceber uma leve tendência a recrutar dentro do grupo, o que é confirmado no caso de membros do sindicato, que apresentam homifilia de 0,401, presente na tabela 2.

A amostra obtida é composta de 264 músicos de jazz de Nova York, em 2001. As características analisadas estão presentes na tabela 1. O programa fornece as estimativas simples da proporção, além da estimativa ajustada pela ponderação do método RDS, a ponderação se dá pelo tamanho médio estimado da rede do grupo e pelo comportamento de recrutamento do grupo, isto é, se algum grupo tem uma capacidade maior ou menor de recrutamento. A proporção de equilíbrio

representa a proporção da amostra no momento em que ela atinge o equilíbrio, mas não leva em conta o tamanho médio da rede dos grupos. Na saída do programa também é obtido um intervalo de confiança de 95% baseado na distribuição gerada via *bootstrap.*

Os resultados foram obtidos utilizando a opção *dual component* com *mean cell size* igual a 12 na estimação do tamanho médio da rede de contatos, pois esta opção produz as estimativas mais estáveis (Heckathorn, 2007). Como este é um exemplo, o número de reamostragens do *bootstrap* que gerou os intervalos de confiança foi deixado no padrão (2500). Para otimizar a precisão, um número de pelo menos 15000 é recomendado, mas quanto maior esse número mais capacidade computacional é exigida, podendo levar mais tempo para obter os resultados.

Definiu-se o nível alfa dos intervalos de confiança em 5% e foi escolhida a opção de *data-smoothing* aumentado no tipo de algoritmo, pois este previne erros de divisão por zero.

Para fazer a análise da variável contínua idade, esta foi classificada em quatro faixas, pois neste método é o único meio de analisar variáveis deste tipo. A divisão delas foi feita de modo que nenhuma categoria ficasse com poucos membros, pois isto costuma resultar em problemas na estimação dos intervalos de confiança.

O programa também oferece a opção de calcular quantas ondas são necessárias para alcançar o equilíbrio da amostra. Para definir quantas ondas são necessárias inicia-se considerando que as sementes são formadas por apenas um tipo, o caso mais extremo, por exemplo, todos do sexo masculino, e então baseado na matriz de transição estimada pela amostra verifica-se as proporções esperadas a cada onda até que não haja variação na proporção dos grupos maior que uma taxa definida. No caso deste exemplo foi deixada a opção padrão, que é 2%.

Foi verificado o número de ondas necessárias para todas as variáveis e partindo de todos os grupos, isto é, no caso do sexo, partia-se tanto de uma semente só de mulheres quanto de uma só de homens para verificar como isso afetava no número de ondas necessárias para se chegar ao equilíbrio.

A variável que chega mais rápido ao equilíbrio é a referente ao músico já ter seu trabalho transmitido no rádio, precisando de apenas uma onda para atingir o equilíbrio, enquanto as variáveis sexo, raça e participação ou não de sindicatos variaram de duas a três ondas para atingir o equilíbrio, sendo o caso da idade o mais complicado e levando de 5 a 6 ondas para se estabilizar.

Tabela 1

Características da amostra de músicos de jazz de Nova York, 2001.

Variável	N = 264*	Bruta**	Equilíbrio	Ajustada	IC 95%
Idade (anos)	263				
≤ 35	65	0,247	0,292	0,380	0,262-0,489
36-45	60	0,228	0,250	0,209	0,151-0,297
46-55	79	0,300	0,272	0,292	0,184-0,375
≥ 56	59	0,224	0,186	0,119	0,078-0,182
Raça	259				
Branca	142	0,548	0,535	0,531	0,425-0,637
Negra	85	0,328	0,341	0,360	0,264-0,468
Outras	32	0,124	0,123	0,109	0,060-0,155
Sexo	259				
Masculino	191	0,737	0,773	0,762	0,656-0,842
Feminino	68	0,263	0,227	0,238	0,158-0,344
Sindicato	263				
Sim	105	0,399	0,413	0,250	0,183-0,324
Não	158	0,601	0,587	0,750	0,676-0,817
Tocou no rádio	253				
Sim	208	0,822	0,817	0,751	0,662-0,849
Não	45	0,178	0,183	0,249	0,151-0,338

* Tamanho total da amostra, ao lado do nome das variáveis aparecem os totais válidos (sem dados faltantes)

** Estimativa simples da proporção populacional (proporção da amostra)

Podemos ver pelas estimativas de proporção da amostra e do equilíbrio que apenas a idade ficou um pouco distante ainda, talvez porque precisasse de mais algumas ondas para chegar ao equilíbrio. Nota-se também que na variável idade, a categoria de músicos mais jovens parece ter sido subestimada enquanto a categoria de músicos mais velhos se faz mais presente na amostra, isso pode ser devido ao fato do maior número de ligações dos mais velhos, como é mostrado na figura 3. Caso parecido com o que ocorre nas variáveis que envolvem a participação no sindicato e tocar no rádio, onde os que responderam positivamente aparecem mais na amostra do que na estimativa da população.

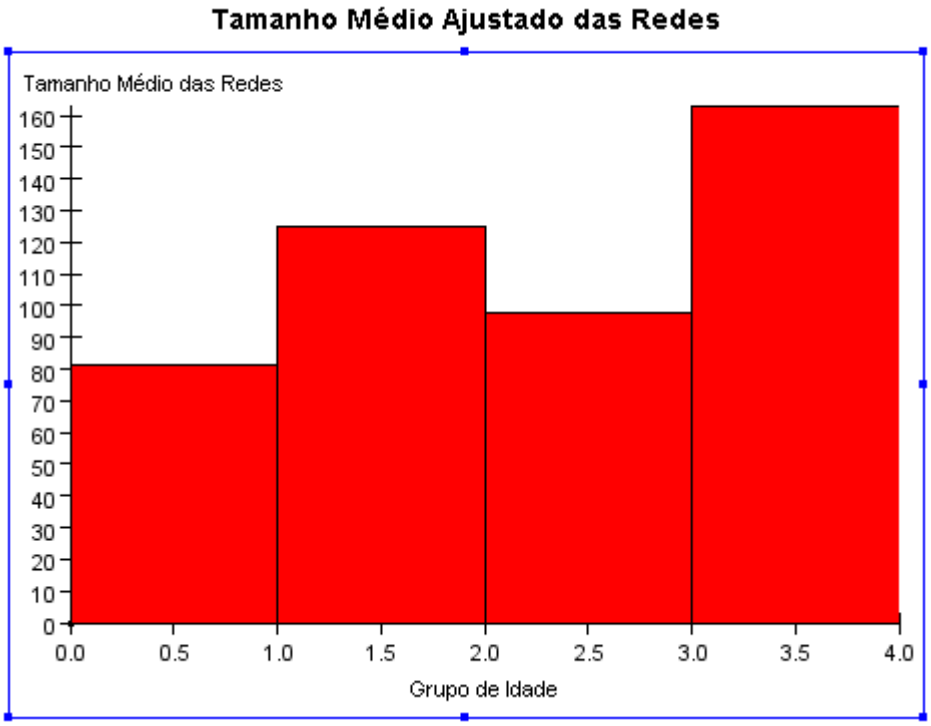


Figura 3: Tamanho médio das redes por categoria de idade

Na tabela 2 encontram-se as estimativas do tamanho médio da rede de contatos em cada grupo e também estimativa da homofilia. Podemos ver que os músicos de jazz mais velhos, os que já tiveram seu trabalho transmitido em alguma rádio e os que fazem parte do sindicato apresentam um número maior de contatos dentro da população, isso justifica o fato de as proporções deles na amostra serem maiores do que realmente o são dentro da população, dada a maior probabilidade deles serem recrutados no processo amostral.

Podemos observar na tabela 2 que há um nível intermediário de homofilia no grupo de músicos homens e no grupo daqueles que participam de sindicato. Isso indica que eles costumam recrutar com mais frequência dentro do próprio grupo e caberia investigar por quais razões isto acontece.

Tabela 2

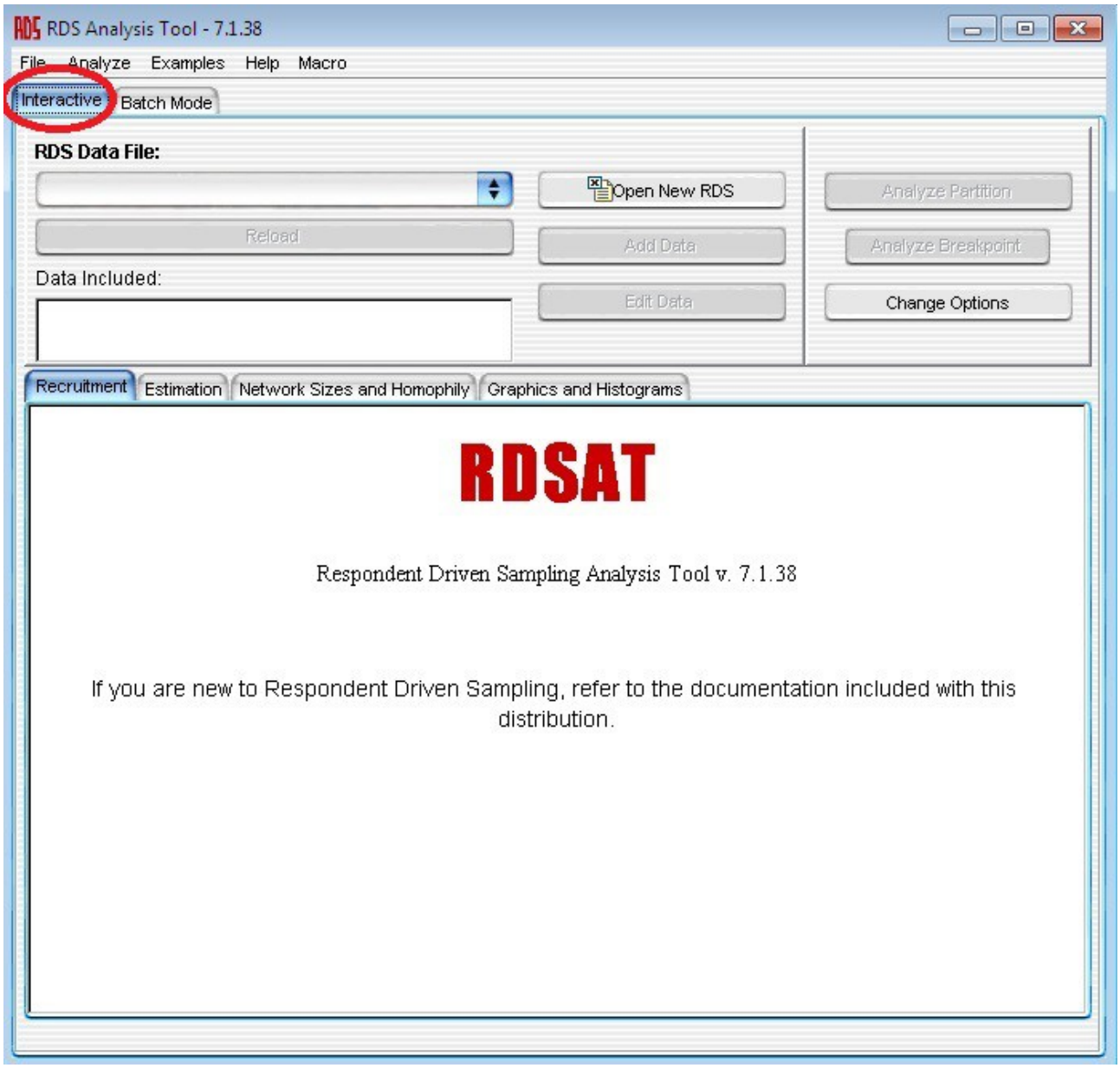
Tamanho médio da rede e homofilia da amostra de músicos de jazz de Nova York, 2001.

Variável	N = 264* (na amostra %)	Média ajustada do tamanho da rede de contatos	Homofilia
Idade (anos)	263 (100)		
≤ 35	65 (24,7)	81,71	0,288
36-45	60 (22,8)	126,79	0,164
46-55	79 (30,0)	98,92	0,081
≥ 56	59 (22,4)	165,60	0,286
Raça	259 (100)		
Branca	142 (54,8)	107,47	0,260
Negra	85 (32,8)	101,08	0,259
Outras	32 (12,4)	120,56	0,073
Sexo	259 (100)		
Masculino	191 (73,7)	110,51	0,309
Feminino	68 (26,3)	103,85	0,265
Sindicato	263 (100)		
Sim	105 (39,9)	180,55	0,411
Não	158 (60,1)	85,54	-0,081
Tocou no rádio	253 (100)		
Sim	208 (82,2)	118,18	0,295
Não	45 (17,8)	79,72	-0,134

* Tamanho total da amostra. Os valores ao lado das variáveis representam os totais válidos (sem missing)

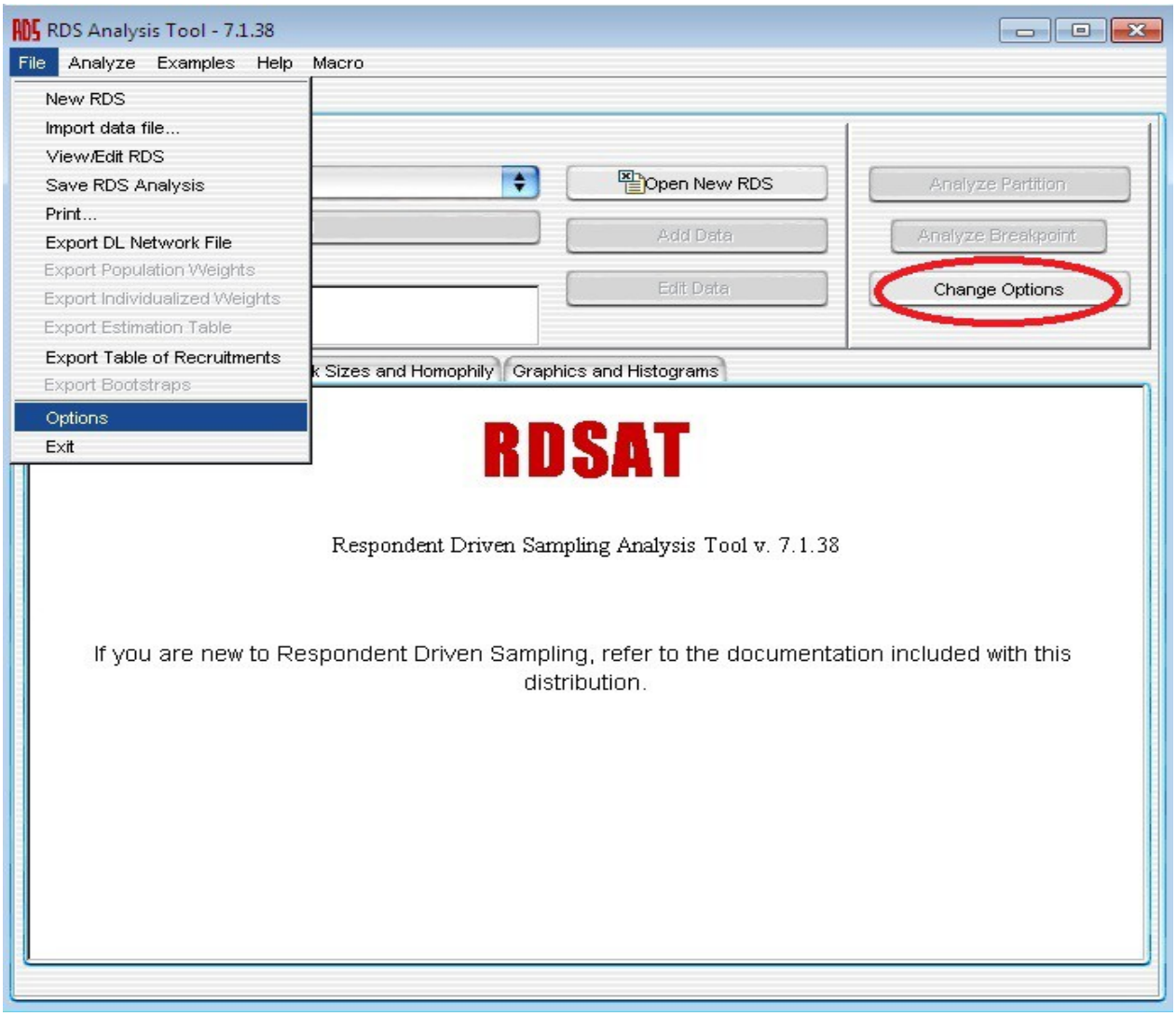
3.12.1) Passo a passo da análise do exemplo

Para demonstrar como se chegou aos resultados obtidos no exemplo, segue um passo a passo básico do programa RDSAT 7.1.

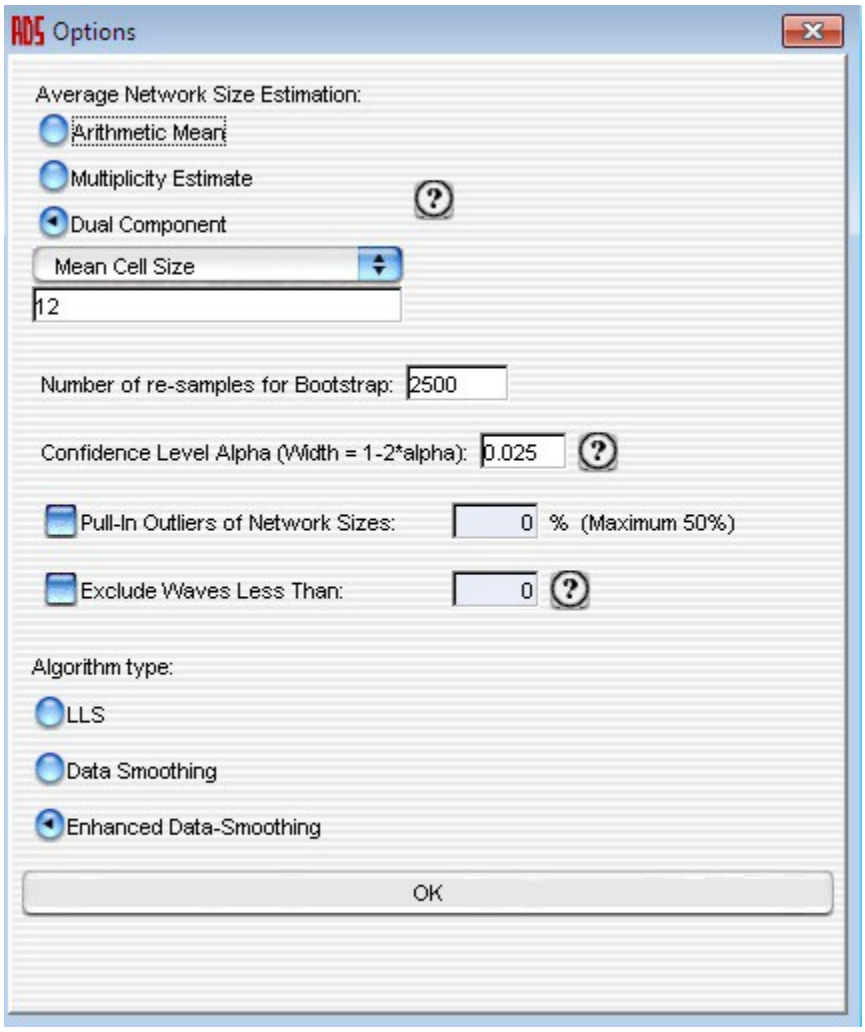


Esta é a tela inicial do programa. Existem dois modos de operação, o *interactive* e o *batch mode*. O primeiro utiliza do "aponte e clique" e serve para analisar um arquivo por ver. O segundo permite ao usuário salvar "trabalhos" e realizar múltiplas análises em um ou mais arquivos.

Para iniciantes é recomendado começar pelo *interactive.* E este será o modo utilizado no exemplo.

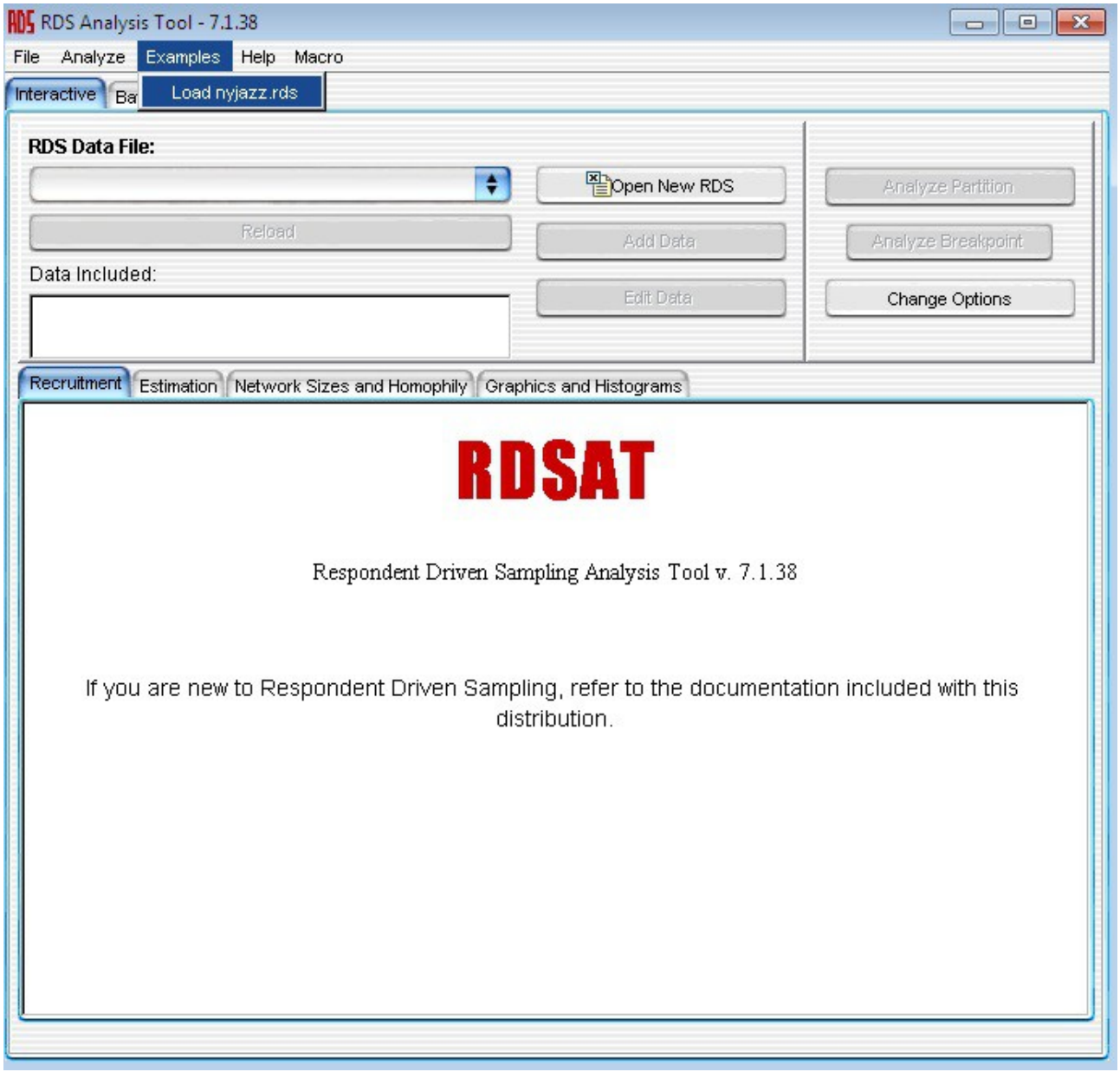


A primeira coisa a se fazer antes da análise é definir as opções. Pode-se perceber já aqui que podemos fazer as coisas por dois caminhos, pela barra de ferramentas ou pelos ícones.

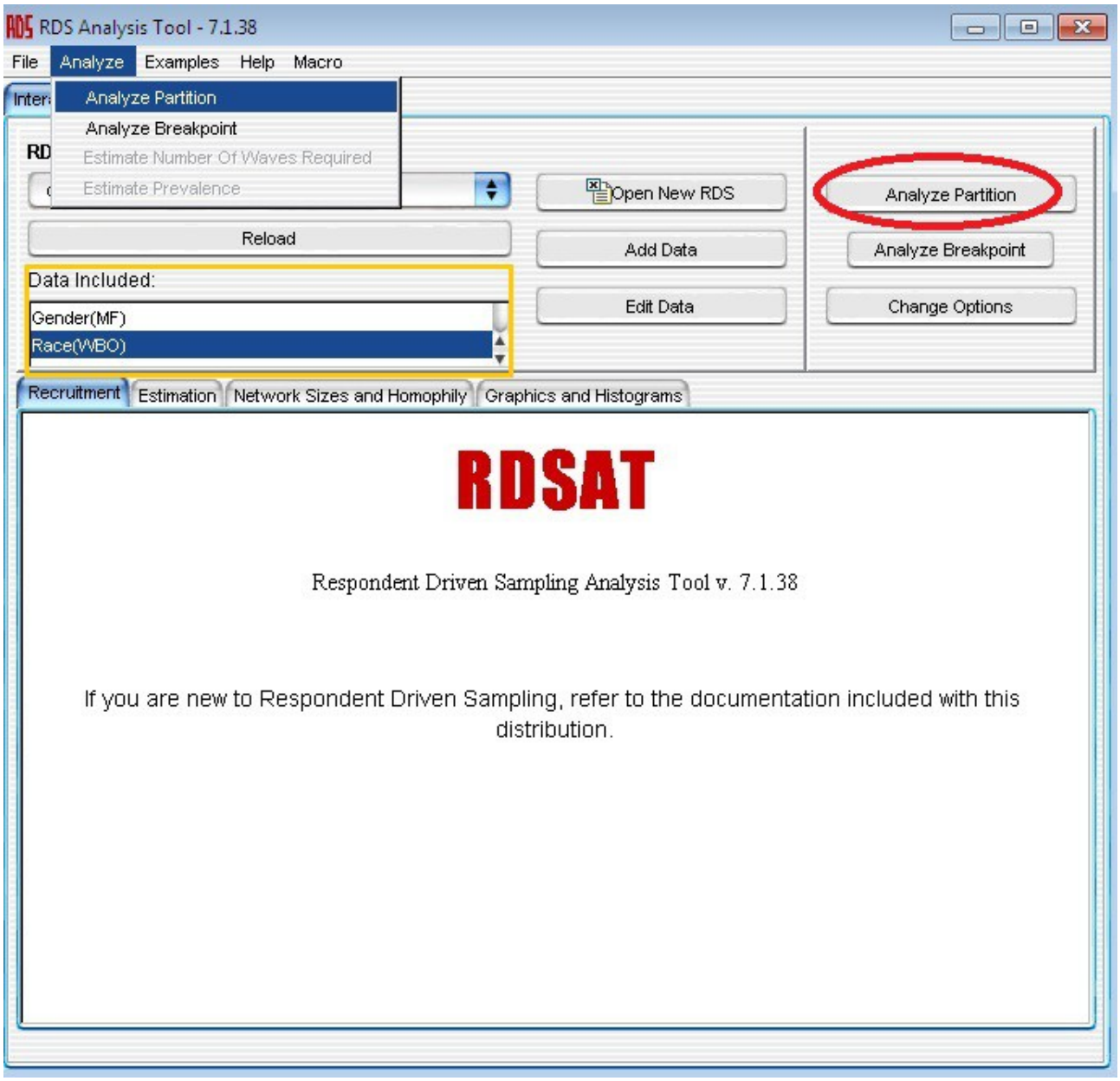


Estas opções serão consideradas ao se realizar a análise. Recomenda-se marcar *dual component* com *mean cell size* = 12 pois esta configuração produz as estimativas mais estáveis.

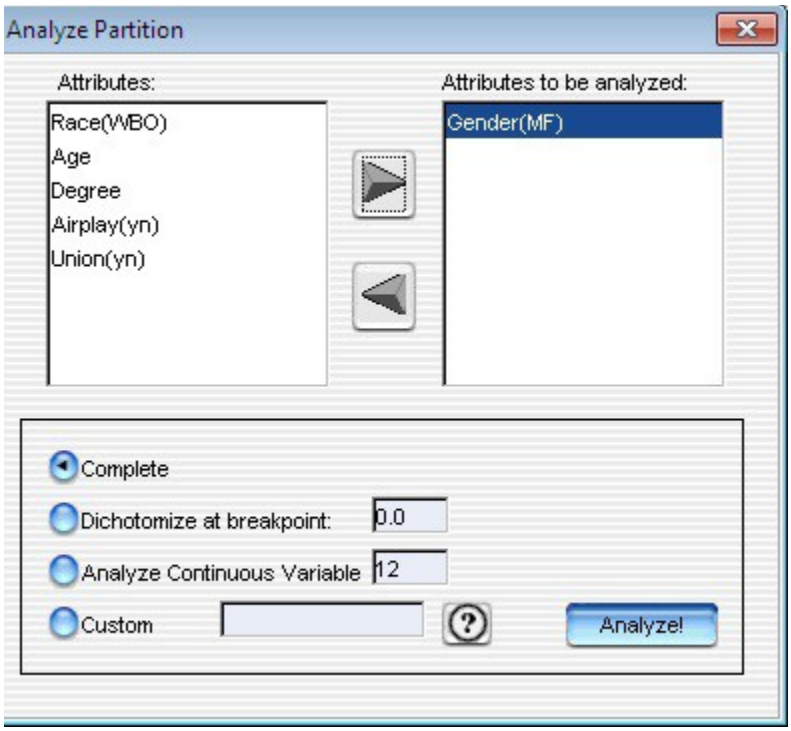
O número de reamostragens para o *bootstrap* precisa ser de no mínimo 2500 para intervalo de confiança precisos, quanto maior esse número mais precisos os intervalos serão, mas levarão mais tempo computacional. O alfa está definido para que intervalos de confiança bilaterais tenham uma confiança de 95%. As opções indica se quer retirar da amostra dados de respondentes cujos tamanho da rede são *outliers*. Definindo-se a percentagem em 5%, por exemplo, excluiria da amostra aqueles com os percentis abaixos de 5 e acima de 95. Recomenda-se que este valor seja pequeno, não ultrapassando os 10%. A próxima opção determina se serão excluídas ondas da amostra, normalmente se deixa desmarcada. Na opção de algortimo recomenda-se deixar marcada a opção *enhanced data-smoothing*, que evita problemas de divisão por zero.



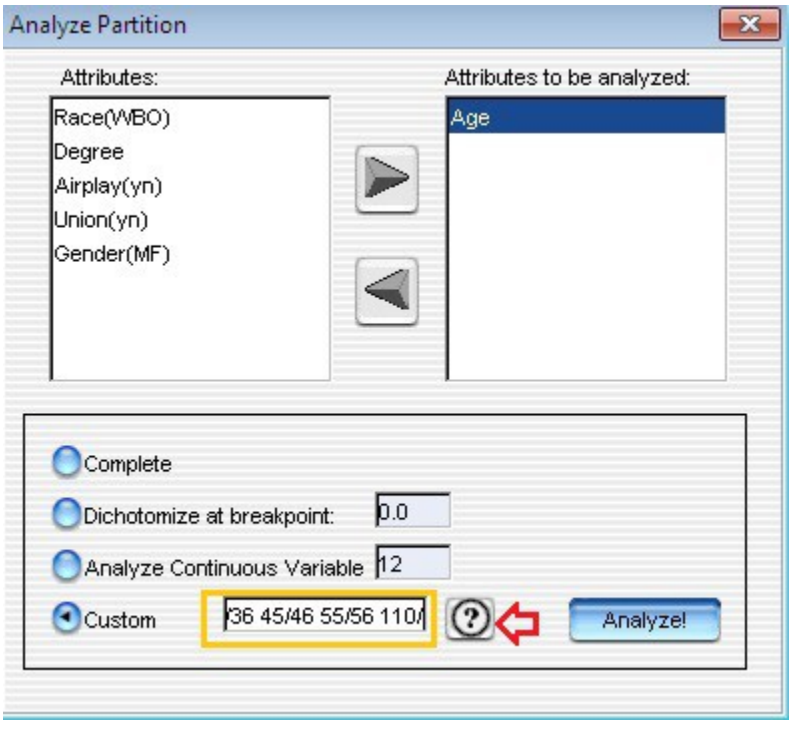
No próximo passo carregamos o arquivo de exemplo presente no programa.



Como podemos ver, agora apareceram as variáveis que podem ser analisadas e foram disponibilizadas novas opções para que se possa fazer analisar os dados. Vamos utilizar a *analyze partition.*

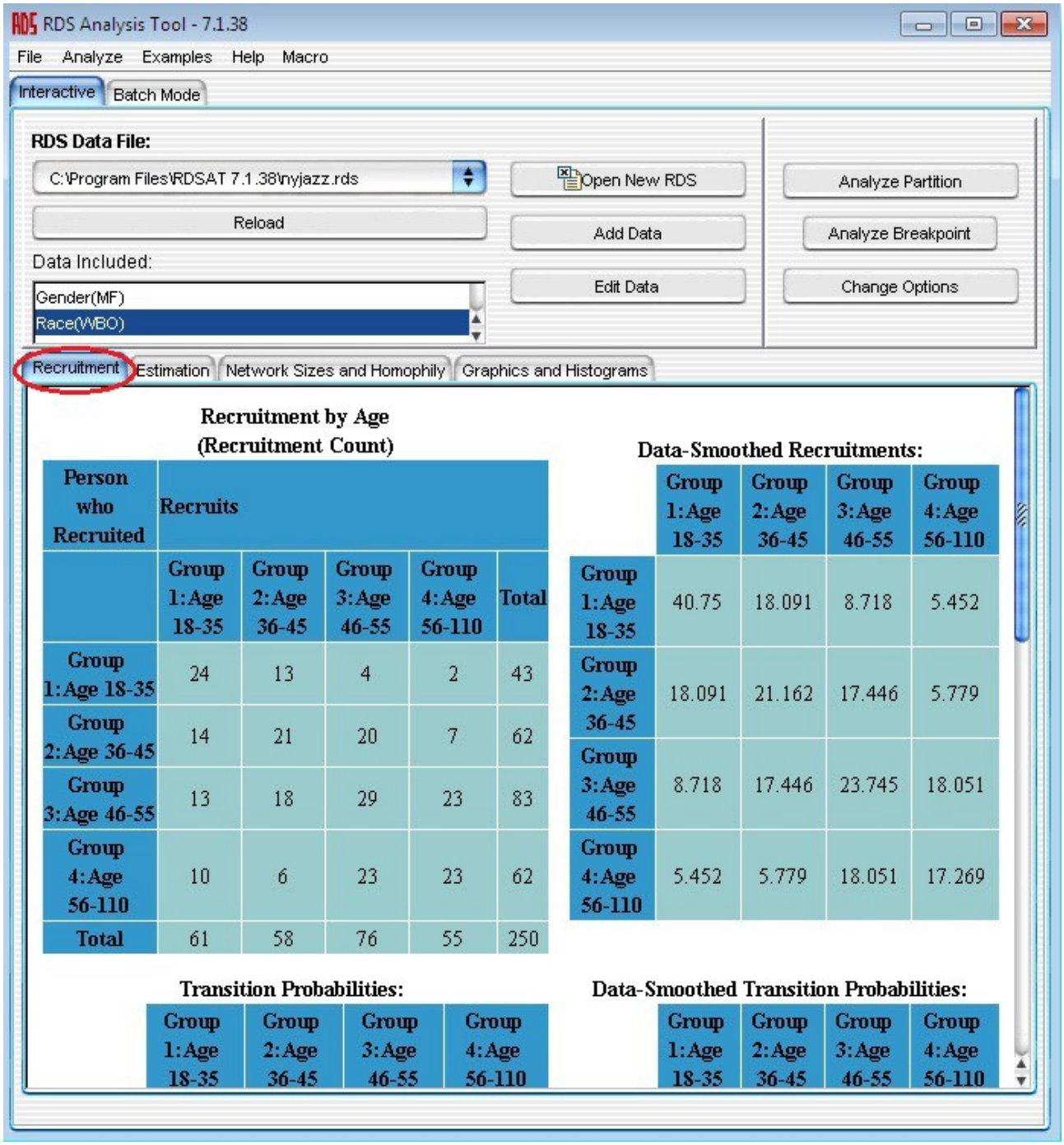


Por exemplo aqui, a variável sexo foi passada para a direita para ser analisada.

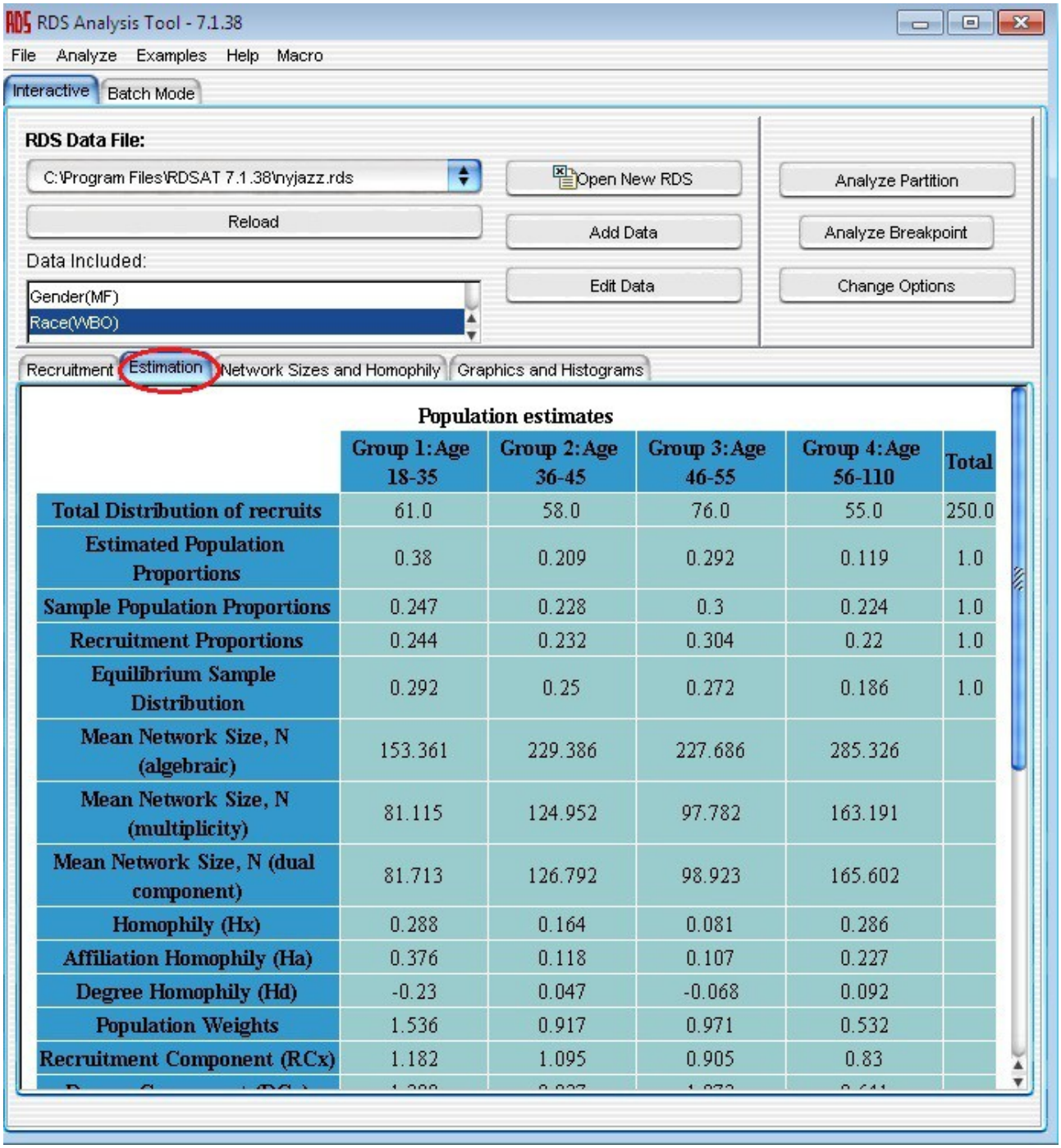


Mas no caso pode ser uma variável contínua como a idade também,então na opção *custom* definem-se as categorias. Para saber como definir as categorias pode-se clicar no ponto de interrogação. Pode-se definir uma categoria indo até infinito mas, apesar do programa oferecer essa

opção, costuma apresentar problemas na formação das categorias. Depos de definidas as categorias, clica-se em *analyze!*.



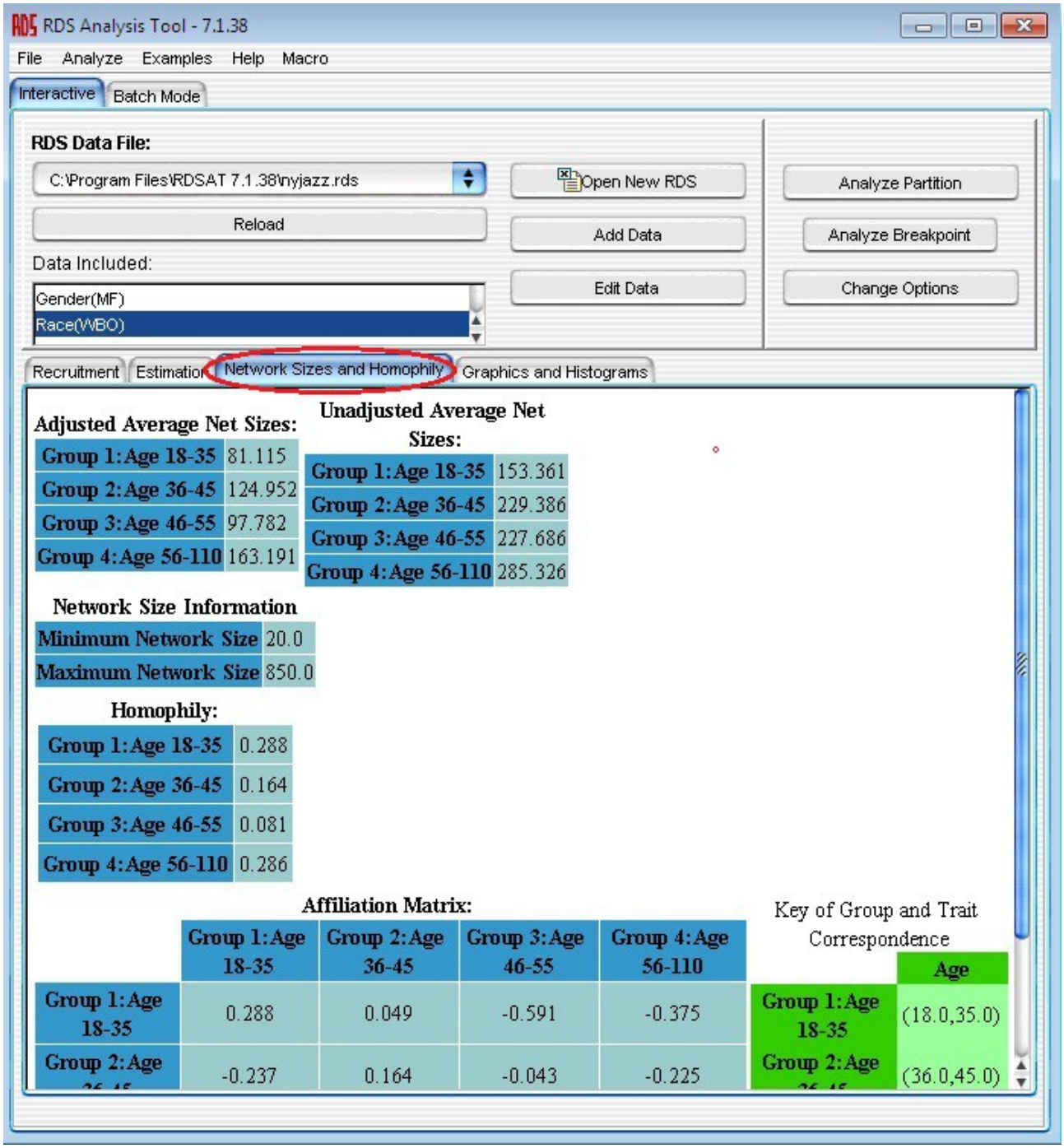
Ao fazer a análise os resultados aparecem na janela maior. Na primeira aba temos os resultados sobre o recrutamento.



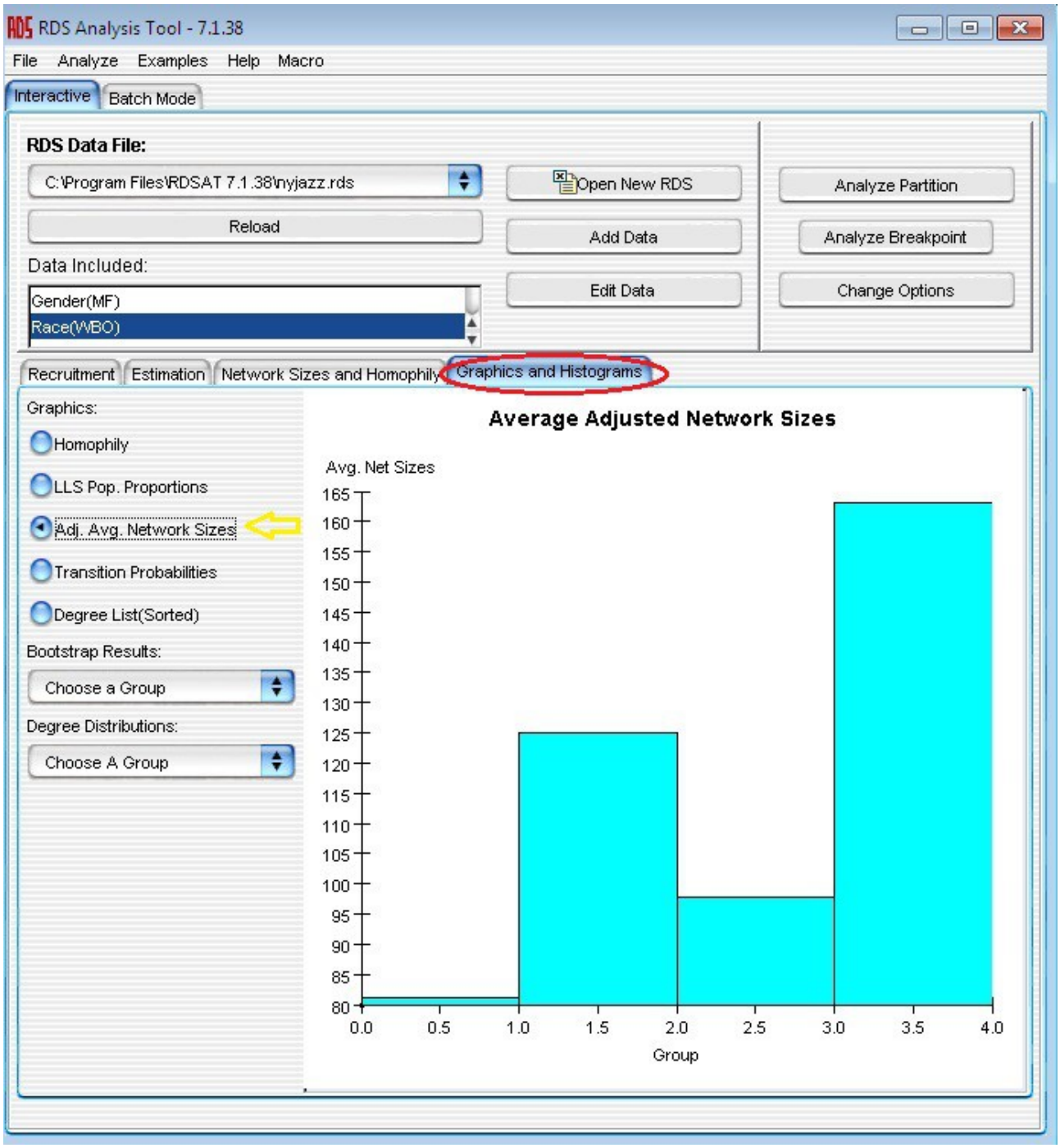
Population estimates

	Group 1:Age 18-35	Group 2:Age 36-45	Group 3:Age 46-55	Group 4:Age 56-110	Total
Total Distribution of recruits	61.0	58.0	76.0	55.0	250.0
Estimated Population Proportions	0.38	0.209	0.292	0.119	1.0
Sample Population Proportions	0.247	0.228	0.3	0.224	1.0
Recruitment Proportions	0.244	0.232	0.304	0.22	1.0
Equilibrium Sample Distribution	0.292	0.25	0.272	0.186	1.0
Mean Network Size, N (algebraic)	153.361	229.386	227.686	285.326	
Mean Network Size, N (multiplicity)	81.115	124.952	97.782	163.191	
Mean Network Size, N (dual component)	81.713	126.792	98.923	165.602	
Homophily (Hx)	0.288	0.164	0.081	0.286	
Affiliation Homophily (Ha)	0.376	0.118	0.107	0.227	
Degree Homophily (Hd)	-0.23	0.047	-0.068	0.092	
Population Weights	1.536	0.917	0.971	0.532	
Recruitment Component (RCx)	1.182	1.095	0.905	0.83	

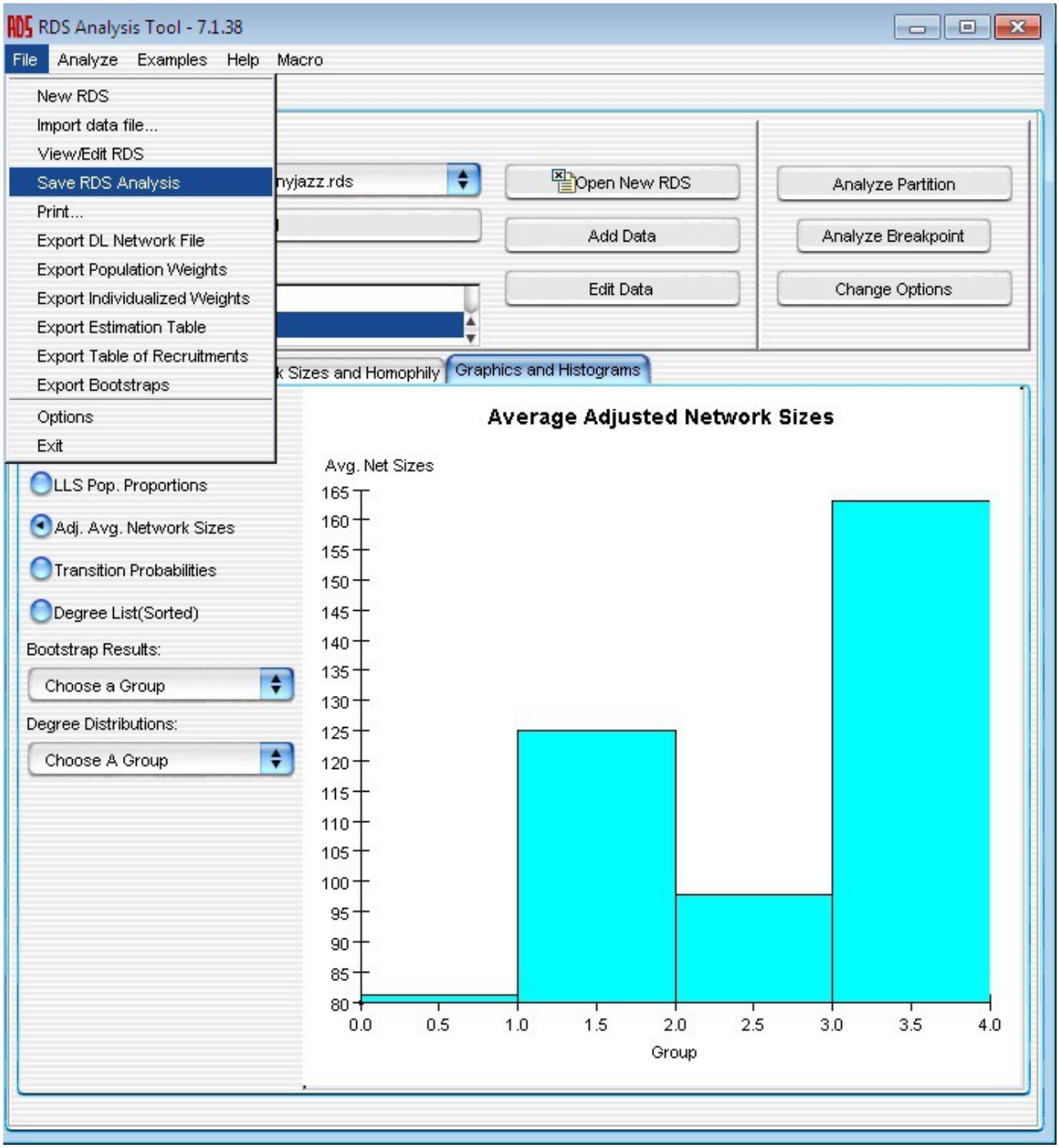
Na segunda aba temos as estimativas de proporção, entre outras medidas.



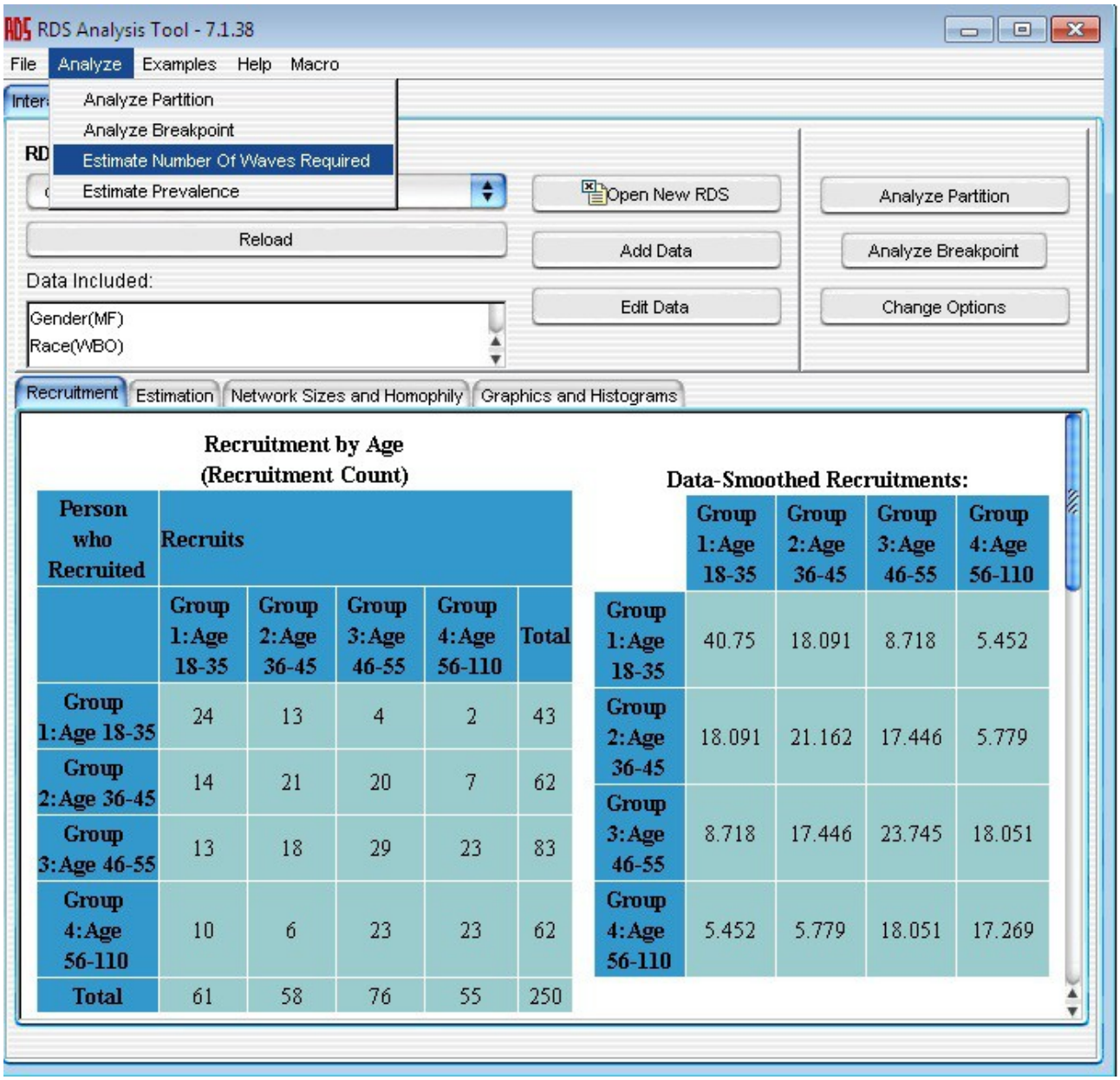
Na terceira aba temos resultados das estimativas dos tamanhos médios das redes, além de estimativas da homofilia.



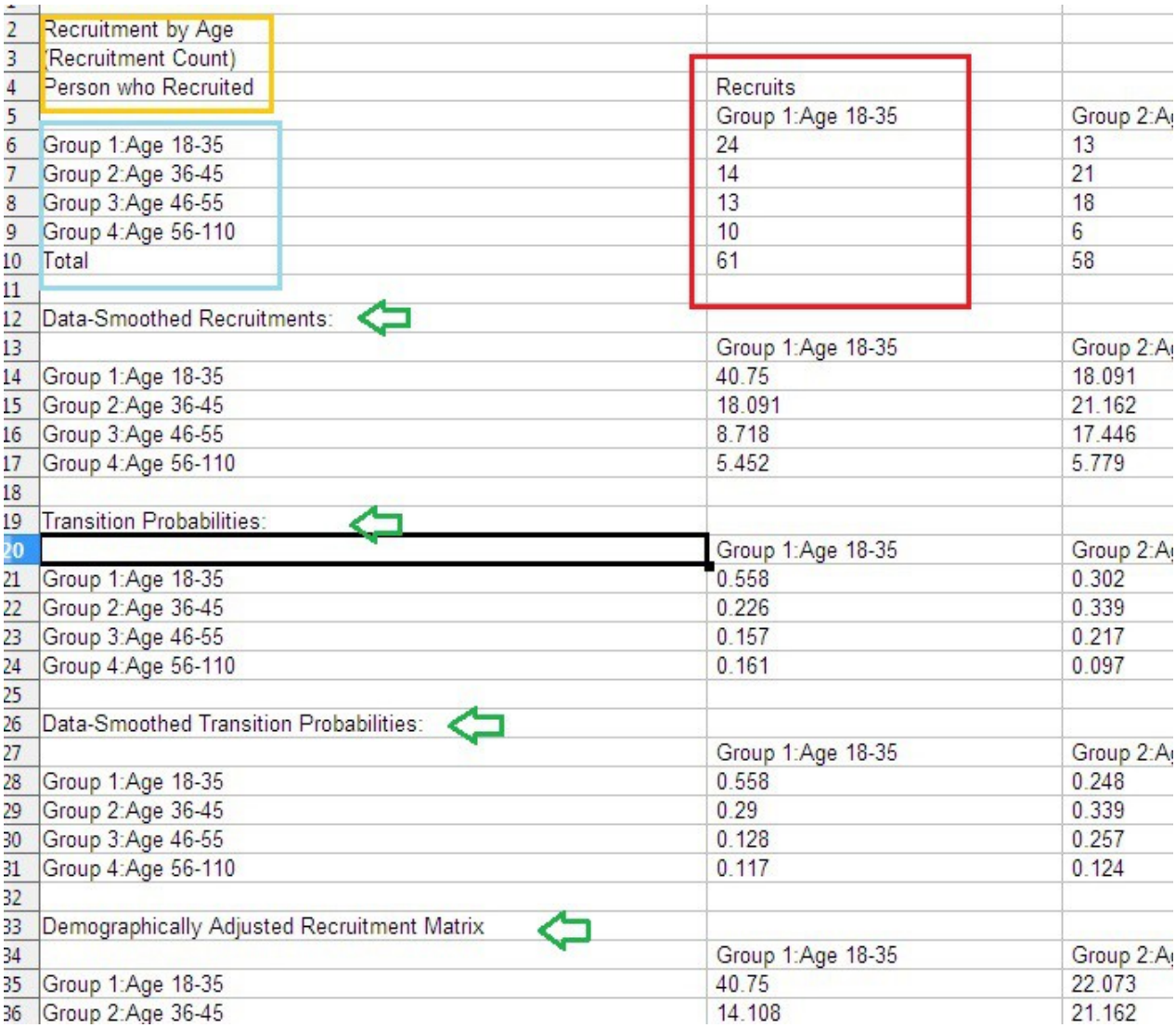
Na quarta aba temos as opções gráficas do programa. Por exemplo, clicou-se onde está indicado para solicitar um gráfico dos tamanhos médios da redes nos grupos de idade. Clicando em cima dos gráficos podemos mudar opções como escala, legendas e cores. Depois de formatar o gráfico pode-se salvá-lo em algum de diversos formatos de imagem disponíveis.



Depois de feita a análise, nesse caso da variável idade, pode-se salvar os resultados em um arquivo.



Além disso, pode-se estimar o número de ondas necessárias para o equilíbrio conforme os dados do recrutamento da amostra.



2	Recruitment by Age		
3	(Recruitment Count)		
4	Person who Recruited	Recruits	
5		Group 1:Age 18-35	Group 2:A
6	Group 1:Age 18-35	24	13
7	Group 2:Age 36-45	14	21
8	Group 3:Age 46-55	13	18
9	Group 4:Age 56-110	10	6
10	Total	61	58
11			
12	Data-Smoothed Recruitments:		
13		Group 1:Age 18-35	Group 2:A
14	Group 1:Age 18-35	40.75	18.091
15	Group 2:Age 36-45	18.091	21.162
16	Group 3:Age 46-55	8.718	17.446
17	Group 4:Age 56-110	5.452	5.779
18			
19	Transition Probabilities:		
20		Group 1:Age 18-35	Group 2:A
21	Group 1:Age 18-35	0.558	0.302
22	Group 2:Age 36-45	0.226	0.339
23	Group 3:Age 46-55	0.157	0.217
24	Group 4:Age 56-110	0.161	0.097
25			
26	Data-Smoothed Transition Probabilities:		
27		Group 1:Age 18-35	Group 2:A
28	Group 1:Age 18-35	0.558	0.248
29	Group 2:Age 36-45	0.29	0.339
30	Group 3:Age 46-55	0.128	0.257
31	Group 4:Age 56-110	0.117	0.124
32			
33	Demographically Adjusted Recruitment Matrix		
34		Group 1:Age 18-35	Group 2:A
35	Group 1:Age 18-35	40.75	22.073
36	Group 2:Age 36-45	14.108	21.162

O arquivo de resultados está no formato *.rds* mas pode ser aberto em uma planilha e tem essa aparência. Os resultados que apareciam separados por abas aparecem aqui todos juntos em uma só planilha, com exceção da parte gráfica.

3.13) Pontos positivos

O método RDS, por ser uma adaptação da estratégia de amostragem em bola de neve, traz consigo os benefícios desse método já citados anteriormente. Além de que resolve alguns problemas na metodologia como a abordagem ética, no modo de recrutamento o pesquisador não recebe o contato do recrutado, mas esse por sua vez vem até o pesquisador após ser recrutado. O método do duplo incentivo diminui problemas como mascaramento e viés de indivíduos mais cooperativos, pois oferece recompensas por recrutamentos feitos, diminuindo a diferença de comportamento no recrutamento entre aqueles mais interessados em cooperar naturalmente e aqueles que o fazem pelo incentivo.

Além disso, sendo satisfeitas as suposições fornece estimativas não-viesadas para população, algo que era impraticável no método em bola de neve. Outro fato importante é a questão de que quando a amostra atinge o equilíbrio, se torna independente da escolha inicial das *sementes*, o que era a principal crítica ao método de amostragem em bola de neve, e assim, consegue uma amostra que pode ser considerada aleatória

Esse método também pode ser usado combinado com outro métodos, por exemplo *network sampling* (Sudman et al, 1988) e investigação etnográfica, pois mesmo partindo de um grupo de *sementes* que pode ser viesado, depois de um certo número de ondas acaba gerando uma amostra não-viesada.

3.14) Problemas

Embora o método RDS tenha tido sucesso em estudar um gama variada de populações escondidas e as estimativas tenham se mostrado não viesadas, tanto analítica quanto computacionalmente, permanece a questão se a teoria e suposições da RDS podem ser aplicadas realisticamente aos dados reais.

Salganik e Heckathorn (2004) mostraram que as estimativas eram não-viesadas, tanto analiticamente quanto computacionalmente, mas descobriu-se mais tarde que a variância dos estimadores era demasiado alta. Sendo assim, mesmo que seja não-viesado, ou seja, na média acerta, a precisão é baixa e não dá aos pesquisadores que só podem fazer uma amostra, nenhuma garantia de que as estimativas estejam sequer próximas da realidade.

Além da variância alta, Verdery et al (2013) relata que há problemas na variância dos estimadores RDS, sendo que estão viesadas no sentido de subestimá-las. O problema é agravado quando a suposição de que o processo amostral se comporte como uma cadeia de Markov de primeira ordem não é válida, o que ocorre com frequência. Os pesquisadores então não podem confiar nos intervalos de confiança fornecidos, pois eles teriam nenhuma validade.

Quando as suposições do processo amostral não são seguidas na prática (recrutamento não-ramificante, declaração precisa do seu grau, amostragem feita com reposição, reciprocidade), mostram aumento no viés para estimadores da média (Gile e Handcock, 2010; Neely, 2009; Lu et al., 2012).

Outra questão é o que se fazer com dados da amostra fora de equilíbrio, que são aqueles recrutamentos feitos nas ondas precedentes à que se atingiu o equilíbrio. Elas não podem ser consideradas não-viesadas porque ainda têm influência da seleção dos respondentes iniciais, mas a retirada deles representaria também numa perda de informação importante e de dados já coletados pelo pesquisador. Wejnert e Heckathorn (2008) acabaram por concluir que incluir os membros pré-equilíbrio na amostra representava pouca diferença e aliviou um pouco essa situação.

Por ser um método recente, ainda há muito a ser desenvolvido quanto a estimadores, e outros parâmetros como coeficientes de correlação e regressão permanecem ainda subdesenvolvidos.

4) CONSIDERAÇÕES FINAIS

O método de amostragem em bola de neve e sua adaptação, RDS, mostram-se efetivos no seu objetivo inicial de alcançar as populações escondidas ou difíceis de encontrar. O método bola de neve vem sendo usado de forma quase que informal pelos pesquisadores, principalmente também no seu interesse de conscientização das populações-alvo quanto à questões de comportamento benéficos e cuidados da saúde. Por sua vez, o método *respondent-driven sampling* traz todos os benefícios da amostragem em bola de neve, sendo rápido de aplicar, barato e de uma necessidade menor de pessoal que outros métodos e tem uma metodologia prática na sua estrutura de distribuição de cupons para o controle da estrutura do recrutamento que permite ao pesquisador um controle maior sobre o processo amostral do que o método amostragem em bola de neve proporciona.

Todos esse benefícios são comprovados pela sua popularidade, principalmente na área da saúde pública, mas também com potencial de atuação nas mais diversas áreas, onde haja os desafios de amostras populações escondidas ou difíceis de encontrar.

Começou em pequenas cidades dos EUA e agora já é aplicada em diversas partes do mundo (Malekinejad, 2008) no monitoramento de comportamento de populações de risco em relação ao HIV, inclusive no Brasil, já tendo havido estudos feitos em homens que fazem sexo com homens em Fortaleza (Kendall et al., 2008) e trabalhadoras do sexo em Santos (Morell et al.,2007).

Apesar disso tudo, na parte da estimação ainda há muito o que se fazer para conseguir obter resultados confiáveis deste tipo de amostragem. Precisa-se mudar algumas suposições que talvez sejam muitos rígidas, além de desenvolver uma metodologia que consiga empregar melhor as informações da amostra para produzir estimadores mais robustos. Por enquanto, apenas seria recomendado aos pesquisadores que utilizam o método RDS que tenham muito cuidado ao interpretar esses resultados, pois o uso descuidado e descompromissado pode levar a conclusões errôneas.

Quanto ao futuro do método há esperança, por se tratar de um desenvolvimento recente, e com o melhoramento da computação e das teorias pode-se chegar a obter estimadores confiáveis para esse tipo de método de amostragem. Há novas abordagens para estimação RDS sendo desenvolvidas (Gile, 2011; Berchenko et al., 2013), além de outros métodos de amostragem envolvendo cadeia de referências (Mouw e Verdery, 2012).

Referências Bibliográficas

ATKINSON, Rowland**;** FLINT**,** John. Acessing hidden and hard-to-reach populations: Snowball research strategies.**Social Research Update.** v.33 Surrey:Department of Sociology, University of Surrey, 2001. p. 1-4

BERCHENKO, Yakir; ROSENBLATT, Jonathan; SIMON, D.W. Frost. Modeling and Analysing Respondent Driven Sampling as a Counting Process. **arXiv:1304.3505**. 2013. Disponível em: <http://arxiv.org/abs/1304.3505>. Acessado em: 20 novembro 2013.

BERG, S. Snowball Sampling**.** Em KOTZ, S.; JOHNSON, N.L. **Encyclopedia of Statistical Sciences**. v. 8, 1988 p. 528-532

BIERNACKI, P.; WALDORF, D. Snowball sampling: Problems and techniques of chain referral sampling. **Sociological Method Research**. v.10, 1981 p. 141-163

BOLLABÁS, Béla. **Random Graphs**. Cambridge: Cambridge University Press, 1985.

BRIGNOL, M.S. Sandra. **Estudo epidemiológico da infecção por HIV entre homens que fazem sexo com homens no município de Salvador.** Tese de Doutorado, Instituto de Saúde Coletiva, Universidade Federal da Bahia, Salvador, 2013 p. 56.

COLEMAN, J.S. Snowball sampling: Problems and techniques of chain referral sampling. **Human Organization.** v.17, 1958 p. 28-36.

FAUGIER, J.; SARGEANT, M. Sampling hard to reach populations. **Journal of Advanced Nursing**. v.26, 1997 p. 790-797

GILE, K.J. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. **Journal of the American Statistical Association**. v.106, 2011. p.135-146.

GILE, K.J.; HANDCOCK, M.S. Respondent-driven sampling: an assessment of current methodology. **Sociological Methodology**. v.40, 2010. p. 285-327

GOODMAN, L.A. Snowball sampling. **The Annals of Mathematical Statistics.** v. 32, 1961. p. 148-170

HECKATHORN, Douglas D. Respondent-driven sampling: a new approach to the study of hidden populations. **Social Problems.** v. 44 , 1997, p. 174-199

HECKATHORN, Douglas D. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. **Social Problems**. v. 49, 2002. p. 11-34

HECKATHORN, Douglas D. Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. **Sociological Methodology**. v. 37, 2007. p. 151-207

HECKATHORN, D. D.; JEFFRI, J. Finding the beat: using respondent-driven sampling to study jazz musicians. **Poetics**. v. 28, 2001. p. 307-329

HECKATHORN, D. D.; JEFFRI, J. Social Networks of Jazz Musicians. Em NATIONAL ENDOWMENT FOR THE ARTS. Changing the Beat: a study of the worklife of jazz musicians, volume 3. **Respondent -driven sampling: survey results by the Research Center for Arts and Culture, Research Division Report 43**. Washington, DC. 2003. p. 48-61

HENDRICKS, V.M.; BLANKEN, P.; ADRIAANS, N.F.P. Snowball sampling: methodological analysis. Em HENDRICKS, V.M.; BLANKEN, P.; ADRIAANS, N.F.P. **Snowball sampling: a pilot study on cocaine use**. Roterdã: IVO, 1992. p. 83-100

HENSLIN, J.M. Studying deviance in four settings: research expericens with cabbies, suicide, drug users, and abortionees. Em DOUGLAS, J. **Research on deviance**. Nova Iorque: Random House, 1972. p. 35-70

KALTON, G. **Introduction to survey sampling**. Beverly Hills, CA: Sage, 1983.

KEMENY, J.G.; SNELL, J.L. **Finite Markov Chains**. Londres: D. Van Nostrand Co. Ltd., 1960.

KENDALL, C. et al. An empirical comparison of respondent-driven sampling, time location sampling, and snowball sampling for behavioral surveillance in men who have sex with men, Fortaleza, Brazil. **AIDS and Behavior**. v. 12, 2008. p. 97-104

KLOVDAHL, A. Urban social networks: some methodological problems and possibilities. Em KOCHEN, M. The small world. Norword, NJ: Ablex Publishing, 1989. p. 176-210

LU, X. et al. The sensivity of respondent-driven sampling. **Journal of the Royal Statistical Society: Series A (Statistics in Society).** v. 175, 2011. p. 191-216

MALEKINEJAD, Mohsen et al. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. **AIDS and Behavior.** v. 12, 2008. p. 105-130

MORELL, Maria G. G. P. et al. Fatores associados à infecção pelo HIV em trabalhadoras do sexo (TS) em Santos-SP. **Saúde Coletiva**. v. 4, 2007.

MOUW, Ted; VERDERY, Ashton M. Network sampling with memory: a proposal for more efficient sampling from social networks. **Sociological Methodology**. v. 42, 2012. p. 206-256

NEELY, W. Whipple. **Bayesian methods for data from respondent-driven sampling**. PhD dissertation, Department of Statistics, University of Winsonsin, Madison, 2009.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Second-generation surveillance for HIV: the next decade. **WHO/CDS/CSR/EDC/2000.5.** Genebra: World Health Organization and UNAIDS Working Group on Global HIV/AIDS and STI Surveillance, 2000.

RAMIREZ-VALLES, Jesus, et al. Confronting stigma: community involvement and psychological well-being among HIV-positive latino gay men. **Hispanic Journal of Behavioral Sciences**. v. 27, 2005. p. 101-119

SADLER, G. R. et al. Research article: recruitment of hard-to-reach population subgroups via adaptations of the snowball sampling strategy. **Nursing & Health Sciences**. v. 12, 2010. p. 369-374

SALGANIK, Matthew J. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. **Journal of Urban Health: Bulletin of the New York Academy of Medicine**. v. 83, 2006. p. 98-112

SALGANIK, Matthew J.; HECKATHORN, Douglas D. Sampling and estimation in hidden populations using respondent-driven sampling. **Sociological Methodology.** v. 34, 2004. p. 193-240

SNIJDERS, T. Estimation on the basis of snowball samples: how to weight. **Bulletin de Méthodologie Sociologique**. v. 36, 1992. p. 59-70

SPREEN, Marinus. Rare populations, hidden populations and link-tracing designs: what and why?. **Bulletin de Méthodologie Sociologique**. v. 36, 1992. p. 34-58

SUDMAN, S.; SIRKEN, M. G.; COWAN, C. D. Sampling rare and elusive populations. **Science.** v. 240, 1988. p. 991-996

VAN METER, K. M. Methodological and design issues: techniques for assessing the representatives of snowball samples. Em LAMBERT, E. Y. **The collection and interpretation of data from hidden populations. NIDA Research Monograph Series.** Rockville, MD: National Institute on Drug Abuse. v. 98, 1990. p. 31-43

VERDERY, Ashton M. et al. Network structure and biased variance estimation in respondent driven sampling. **arXiv:1309.5109.** 2013. Disponível em <http://arxiv.org/abs/1309.5109> . Acesso em: 20 novembro 2013

VOLZ, Erik; HECKATHORN, Douglas D. Probability based estimation theory for respondent-driven sampling. **Journal of Official Statistics.** v. 24, 2008. p. 79-97

VOLZ, Erik et al. **Respondent-Driven Sampling Analysis Tool (RDSAT) Versão 7.1**. Ithaca, NY: Cornell University, 2012.

WATTS, Duncan J.; STROGATZ, Steven H. Collective dynamics of "small world" networks. **Nature**. v. 393, 1998. p. 440-442

WATTERS, John K.; CHENG, Yu-Teh. HIV-1 infection and risk among intravenous drug users in San Francisco: preliminary results and implications. **Contemporary Drug Problems**. v. 14, 1987.

WEJNERT, Cyprian; HECKATHORN, Douglas D. Web-based network sampling: efficiency and efficancy of respondent-driven sampling for online research. **Sociological Methods & Research**. v. 37, 2008. p. 105-134