**What is Respondent Driven Sampling (RDS) and why should we use it?**

*Respondent Driven Sampling* (RDS) is a systematic method of recruiting people to participate in a survey or an interview which produces a “representative sample” of a population when the precise number of people that belong to the group is not known. Surveys are familiar to most people that live in the United States, but they are often done in precise ways by statisticians. For example, in a presidential election year, the public is bombarded with information which tells them that a survey of a “random sample of voters” has found one candidate is ahead of another by “plus or minus” a number of percentage points. Most of the surveys that attempt to predict presidential elections do so by phoning “randomly selected” people and asking them a series of questions about their preferences. Surprisingly, the number of people that researchers need to call in order to make a prediction that is fairly accurate, “plus or minus” a few percentage points, is quite small; often about 1,000 people. But how can it be that for a nation that has more than 300 million people, a survey of just 1,000 people can accurately predict the outcome of a nationwide election?

The short answer to that question is that researchers rely on “random” selection to get a “representative” sample of voters, and that sample is a fairly accurate reflection of the entire population. Random means that, in theory, every voter in the country has an equal chance to be called. Of course, not everyone has a phone at home, especially in the age of cell phones, but survey researchers believe that despite those problems, they can find enough randomly selected people that do have phones to compensate for the ones that might otherwise be missed. Clearly, getting a sample that is representative of the larger population is very important if a survey is to have any value, and random selection is widely thought to be the gold standard for recruiting such a sample of people.

The method of random selection of people works very well when there is a phone book or a known list from which researchers can choose names or numbers to call, but when there is no phone book or list of people from which to choose, random selection is not possible. In short, if you don’t know how many people are in the pool of people that you wish to survey, then you cannot be sure that everyone has an equal chance of being selected, and thus, you cannot be sure that you have a representative sample. This is especially true if the people that you want to survey are hard to find or don’t want to be found, like illegal immigrants or criminals. Clearly, a phone survey is not the best approach to learn about these kinds of people. But if it is not possible to achieve the “gold standard” of survey research – that is, randomly selecting from a known number of people to get a representative sample – are there other ways to get representatives sample of certain types of people? The answer is a qualified yes.

Respondent Driven Sampling ([www.respondentdrivensampling.org](http://www.respondentdrivensampling.org)) (RDS) relies on the same principle that telephone researchers use – that everyone has an equal chance of being recruited into the study – but because it is impossible for the researchers to recruit a random sample from a known number of people, they rely on the people who belong to specific groups to recruit their friends and associates who are like themselves into the study in a systematic way – using numbered coupons that allow us to track how the recruits are connected to each other – that eventually produces a representative sample once a sufficient number of people have been recruited. Here’s how it works.

Starting with the first person recruited into the study, called a “seed,” researchers ask he/she to find 3 people who are “like” themselves and give each of them a numbered coupon inviting them to be in the study. From this seed, the researchers will grow an entire tree populated by people that branch out in myriad directions from the initial seed.

In theory, that first person, the seed, can recruit anyone who is a member of the group that the researchers wish to learn about, but of course, they are most likely to recruit those people that are their friends, relatives or those who are closest to them that fit the study’s description. But wait, you might ask, won’t this method of recruitment produce a biased sample? Won’t people just bring in their friends, thereby missing a lot of people that might be important to include if you expect to end up with a representative sample?

The short answer is yes, they will often recruit their friends and others just like themselves, and the degree to which they do so could bias the sample. For example, a white man may be inclined to recruit other white men, even though a man of any race/ethnicity might qualify for the study. Statisticians call the tendency to recruit people “just like” themselves “homophily” (which means, literally, “love of the same”), and they measure how strong or weak that trend is among the people that participate in research studies to gauge how much impact homophily has on the final sample, that is, to help determine whether the sample is representative of the larger group or not. Even though homophily can be a problem if it skews the sample too much, RDS researchers have shown that if you recruit enough people in a systematic way, the problem can be overcome, and rather quickly at that. The hypothetical example below illustrates why this is so.

Let us imagine that we would like to do a study of people who play on basketball teams in New York City. There is no roster of players anywhere in the city from which we may select a random sample, and we have no idea about how many basketball players might exist, so we have decided to use RDS to recruit a representative sample of them, and we have selected a large public park in the heart of New York City where there are multiple courts that serve thousands of basketball players to begin to recruit a representative sample.

Most of us would agree that people of all stripe play basketball, but they are often segregated into groups when they do so: men tend to play with other men, women play with women, old men play with old men, white men often play with other white men, black men often play with other black men, and so on. Clearly, a representative sample of basketball players would include all of these various kinds of people that play basketball. But if we begin our RDS study of basketball players with a single seed from one of these groups – say, white female basketball players – and we may find that white female basketball players are more likely to recruit other white female basketball players (especially the other players on their team and the players from their opposing team), then it seems logical to conclude that our sample would be biased toward white females. But while that might appear to make sense, the logic and the math of the recruitment strategy that is at the core of RDS recruitment ensures that this bias is quickly overcome. Here’s why.

Below, a small portion of that large public park is illustrated where there are several basketball courts that currently feature 6 teams playing: Red Stars vs. Blue Stars, Yellow Suns vs. Green Suns, and Red Thunder vs. Blue Thunder.

Let us imagine that players on both Stars teams are white females and we choose a player on the Red Star team as our seed:

After we interview her about her involvement with playing the sport, we ask her to recruit any 3 other basketball players in the entire park to participate in the study. Since we know that white females are likely to recruit other white females, we can assume that she will probably recruit 3 females that play on her 5-person team, also white females. Since there are 4 other players on her team, she will choose 3 of them and ask them to participate in the study. Those 3 recruits become the first “wave” of recruits in the study.

Wave 1 Recruits

Teammate 1

Seed

Teammate 2

Teammate 3

After each of those three white females that form Wave 1 show up to get interviewed, they are each asked to recruit 3 more basketball players, a total of 9 more basketball players that would form the 2nd Wave of recruits. Given that there is only 1 more player from the Red Stars that might be recruited into the study, the first wave of recruits must seek additional participants from outside their team, and the team most likely to be recruited are those that form their opposing team of white females on the court, the Blue Stars. But even if the first wave of basketball players recruit all of the members of the opposing team, the Blue Stars, there are still not enough Stars to satisfy the needs of the study. So, some members of the Red Stars in Wave 1 must recruit outside of their team and their opponent’s team to find additional basketball players. The illustration below shows how that process might work.

Wave 4:

81 players

Wave 3:

27 players

Wave 2:

9 players

Wave 1:

3 players

Seed

As the diagram illustrates, Wave 1 Recruits quickly run out of teammates to recruit and they must seek members of other teams to be in the study. If we imagine that the Blue Thunder and the Red Thunder basketball teams that are composed of Black women, or even if one of those teams has a single Black woman as a player, then the sample will quickly expand beyond the province of white women. Clearly, by the time that the RDS project reaches the 4th Wave of recruits and 81 players are needed, then most players are likely to find that many of their friends and co-competitors have already been recruited into the study, and they must seek people from other teams that they do not know very well. At this point, the issue of friends recruiting friends in a way that biases the study becomes much less of a concern. By the time that the 5th Wave is completed, all bias has been erased from the recruitment process and a representative sample has been achieved. Indeed, statisticians have calculated that the biases introduced by problems associated with homophily can be overcome in as few as 5 waves of recruitment, but clearly, longer chains of recruitment will produce more confidence that a representative sample has been achieved. As a rule of thumb, the fewest number of people that an RDS study should recruit to be confident that they have drawn a representative sample is 200 people, but by way of contrast, the Centers for Disease Control (CDC), which uses RDS to learn about the prevalence and spread of AIDS in particular populations, like drug injectors, recruits samples of 500 people in each city they investigate.