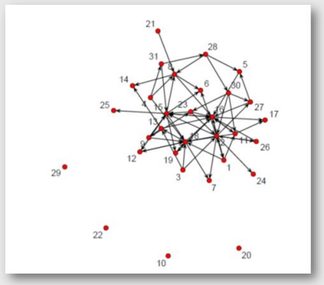
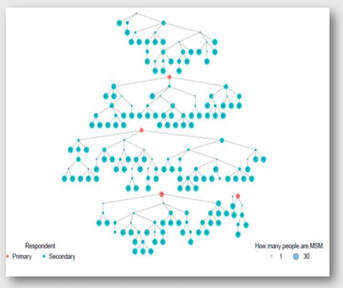
Testing for Preferential Recruitment In RDS

Abhijit Pawar

Department of Mathematics and Statistics, University of Massachusetts Amherst

Advisor: Professor Krista Gile





Diagrams 1&2: Directed Neighborhood Network(left) and Recruitment Tree(Neighborhoods)(right)

What is RDS?

1. Respondent Driven Sampling(RDS) is a sampling method used to measure hard-to-reach populations of individuals with certain characteristics.
2. Hard-to-reach populations are classified as populations of people who are small relative to the general population, and for whom there is little to none data available.
3. In an RDS survey, a person with the particular characteristic is sampled and is then asked about his/her degree, or how people he/she knows who has the same characteristic.
4. Degree- Number of people a respondent knows in the hard-to-reach population.

Data:

1. Dataset was from Kiev which screened the population of people who are MSM(Men who have sex with Men).
2. Each respondent was asked
   a. His degree(how people he knows who are MSM and live in Kiev)
   b. Which Neighborhood of Kiev he lives in(1-31).
   c. Who recruited them for the survey based on the coupon number they had
   d. How many people they recruited for the survey, based on coupon number.
3. There were 200 respondents in the dataset, but only 84 of them gave information on their degree and neighborhood.

Objective:

- To check for any significant preferential recruitment in the dataset in terms of a respondent's neighborhood, district, and proximity(distance) to another respondent's neighborhood
- In order to find out, I conducted three different permutations tests(Using R code) in which I:
  - Found the test statistic from the data(test_stat)
  - Conducted a simulation for 1000 trials in which I got sim_stat
  - Found the p-value= sum(sim_stat<test_stat)/10000, from I determined if there was preferential recruitment

Permutation Test Results

1. Neighborhood/District Permutation Test:

- **Purpose: to test whether or not respondent had a tendency to recruit respondents in the same neighborhood/district as them.**
  - **Steps:**

1.) Test statistic: the number of times a respondent recruited another respondent from his own neighborhood/district, directly from dataset.

- **For Neighborhoods: Test statistic= 6, which means that there were 6 instances of respondents recruiting from their own neighborhood in the dataset.**
- **For Districts: Test statistic= 17, which means that there were 17 instances of respondents recruiting respondents from their own district in the dataset.**

2.) Simulation: For the simulation I generated a random sample of neighborhood/district recruitment ties for 10000 trials from the dataset and test statistic for each trial:

a. After that I generated the frequency of each test statistic in the simulation and created a histogram of them, with the red line as the test statistic

b. From the frequencies I got the p-value:

- **For Neighborhoods: p-value= sum(sim_stats<6)/10000= 0.5655**
- **For Districts: p-value= sum(sim_stats)<17)/10000=0.5751**

c. Both of these p-values indicate that there wasn't significant preferential recruitment for respondents who lived in the neighborhood/district

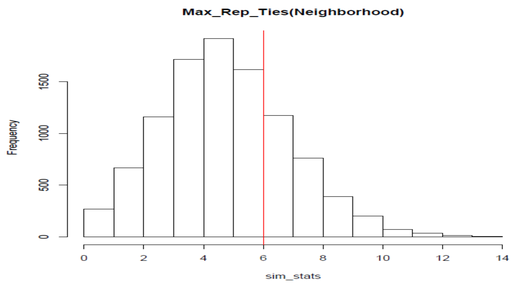


Figure 1: Neighborhood Permutation Test Histogram

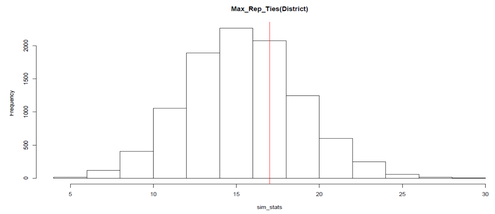


Figure 2: District Permutation Test Histogram

2. Neighborhood Distance Permutation Test:

- **Purpose: to test whether or not respondents had a tendency to recruitment respondents that lived close to their neighborhoods.**
  - **Steps:**
    - **Created a distance matrix whose entries were the distances between all the neighborhoods of Kiev, regardless whether or not they had a recruitment tie in the dataset.**
    - **Test statistic: The sum of all the distances between all the recruitment pairs in the dataset.**
      - **Test statistic= 901.53 km, which is the sum of all distances of the recruitment ties.**

- **Simulation: For the simulation I generated a random sample of neighborhood recruitment ties for 10000 trials from the dataset:**

a. From the simulated ties I used the test statistic function to sum up the distances of the simulated ties(sim_stat):

i. The simulated ties got their distance from the distance matrix and I added up these distances to get the test statistic for each trial.

b. After that I generated the frequency of each test statistic in the simulation and created a histogram of them

c. From the frequencies I got the p-value: p-value=sum(sim_stats<901.53)/10000 which was 0.092, which means that there wasn't significant preferential recruitment for respondents who lived near each other.

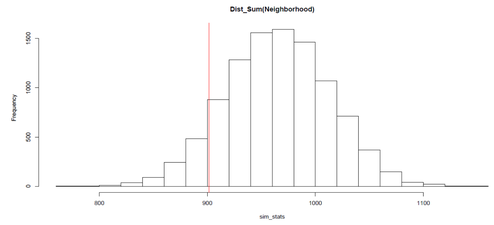


Figure 3: Histogram for Neighborhood Distance Permutation

Conclusions

The results of all the permutations indicate that there wasn't any significant preferential recruitment in the dataset. These results make sense since there wasn't a dominant recruitment relation amongst the neighborhoods. Even though the test statistic for districts was higher for the neighborhoods, it still yielded a high p-value, which indicates that even the district recruitment relation weren't stronger than neighborhood recruitment relations. Even though the neighborhood distance test had the lowest p-value, it wasn't significant enough to show that there was a strong preference for recruiting respondents that lived close to a respondent's neighborhood. But nonetheless, comparing the p-values it is seen that respondents had a higher preference recruiting respondents that lived close to them rather than in the same neighborhood/ district as them. These results aren't surprising since more half the data was missing, therefore any conclusions about the recruitment relation would have been tentative.

Acknowledgments:

Thank you to my advisor Professor Krista Gile for helping me and providing me with guidance through my project I would also like to thank Lisa Johnston for providing me the RDS dataset for Kiev. And Miles Ott(Brown) for his useful comments and suggestions for my permutation tests. As well as Isabelle Beaudry(UMass Amherst) and Kerry Ann Spitzer(MIT), for their feedback on my poster. Finally I would like to thank Mrs. Joanna Barksdale for her generous donations for my REU program.