Universiteit Utrecht

Non-Uniform Sampling in Respondent Driven Sampling (RDS)

Bachelor Thesis

Jaap Nieuwenhuizen, 5966213

Wiskunde

Supervisor:

Dr. Martin Bootsma
Mathematisch Instituut

Bachelor Thesis Mathematics (TWIN)

Non-Uniform Sampling in Respondent Driven Sampling (RDS)

Jaap Nieuwenhuizen
Studentnummer: 5966213
Universiteit Utrecht

June 14, 2019

Abstract

In this research, tweaks on RDS (Respondent Driven Sampling) are being covered. Ultimately, we are interested in reducing the variance of the prevalence. However, to this end, the estimators for the degree and contact type probability are also discussed. Two different sampling methods are being considered: uniform sampling and non-uniform sampling. For the latter, a difference can be made between perfect contact information and imperfect contact information. The use of contact information is paramount for the procedure of non-uniform sampling. And therefore, an analysis of the variance reduction of using contact information has also been investigated. This includes a simple example of a variance comparison simulation between full contact information and exclusively index case information. Included is also a description and discussion of the effects of non-uniform sampling on the attained sample, and the consequences thereof on the variance of the prevalence estimator.

Contents

1 Introduction

Respondent Driven Sampling (RDS) is a method to acquire data from hard-to-reach populations, making use of the snowball or chain referral sampling. Through sampling in general, information from a representative sample is collected, and inferences about the entire population are made. As it is often difficult to find and contact people from the general population with (and potentially due to) certain characteristics of interest, RDS has become increasingly prevalent in gathering data from these types of people. For instance, because these groups are at certain health risks or participate in risky behaviour. Due to being stigmatised or otherwise, a couple of identified examples of such groups are non-social drug users, hermits, homeless people, migrants, illiterates, faith based communities, sex workers, males whom have intercourse with other males, and even jazz-musicians ([5, 3]). Correspondingly, research to the RDS method itself has also continued progressing. Researchers in the field are attempting to find ways to estimate parameters of interest more efficiently and with less uncertainty.

Traditionally, RDS leaves the initiative of recruiting new participants up to the index case. This is thought of to be uniform sampling. Alternatively, one could apply non-uniform sampling, which would imply selecting by certain traits or types of interest; essentially, then, the researcher could choose the new recipients of invitations to participate in the study from the list of contacts of the index case under investigation. Recruiting people often requires coupons or some form of monetary reward, so it is preferable to gather as much information with as little resources.

By using information about this network of interconnected people and their respective traits, one can derive asymptotically proper, unbiased estimators. Given this information about the respective sample, using Markov chain theory, we aim to obtain estimators of prevalence, degree and type homophily of contacts.

Roughly speaking, the degree indicates the number of contacts an index case has, and the homophily is a parameter which indicates a preference for contacts of a similar type as the index case. These are relevant parameters to know and understand when preparing a research which involves inquiring people.

In this paper we suggest two more tweaks to the existing body of measures to improve RDS. We investigate the influence of non-uniform sampling on statistical sampling power, and we further suggest using the full information of every index case in order to gather more data. The benefits of such tweaks could potentially lead to a reduction in time and money, as fewer participants might be needed in order to gain satisfying results. Such recruitment processes could even be simulated. Additionally and importantly, we investigate the accuracy or correlation between estimates of index cases about their contacts and the respective certainty in estimates about prevalence of having the trait of interest and similar statistical estimates, when information about contacts of index cases is retrieved instead of exclusively information about index cases. The issue is that index cases can't necessarily claim with absolute certainty whether their contacts have a certain trait. An income class is for example harder to estimate accurately than gender.

To motivate the potential of non-uniform sampling, we emphasize that initial sampled populations tend not to be representative for an entire population. Through a biased initial sample frame, researchers using RDS intend to collect data from contacts of research participants. A major obstacle in gaining data is another form of bias; the referrals of index cases could plausibly be similar in terms of characteristics as the respective index case. RDS can compensate for such biases. However, RDS can also be further optimalised.

2 Theory

2.1 Assumptions

Before we continue with some underlying theory of the terms and topics which we will present results for, we will make all of our assumptions explicit. These assumptions are important, as they might not represent real world situations. However, they are convenient to produce a tractable mathematical structure.

- The entire network of individuals the researcher is interested in is connected; the population is linked by contacts. This means that everyone knows at least one other person. In fact, we will further make the assumption that the number of coupons (invitations to participate in the research) is always lower than the number of contacts a person has, i.e., all coupons can be distributed.
- Incentives are used to get all sorts of people to cooperate in the form of coupons. Therefore, there is no bias towards more cooperative people in the population to participate in the research.
- Reciprocity in relationships. This means that the contacts of the index cases also have the index case as a contact. Furthermore, the index case knows all contacts, and the respective contacts know the index case. They can identify each other, and estimate whether the respective person has the trait (that is, being part of a certain population) or not.
- The network is assumed to have little structure. Therefore, transitivity in contacts is negligible. So the model is locally tree-like. In other words, very tight interconnected groups of friends are not assumed to occur. This makes for a so-called configurational model. There is some structure, as the model can allow a varying degree, and crucially there can be a preference for contacts with certain traits.
- It is also assumed that the population at large is infinitely big (or at least sufficiently large). This assumption implies that new recruits can be drawn with replacement, as it is assumed that the chance of recruiting the same person is negligible. So, essentially, the state of the system is only determined by the last recruit. Therefore, a Markov process is applicable.
- We assume the chain does not break. So, every coupon successfully recruits a new participant.

2.1.1 Degree Distribution

We aim to show that the chance to receive a coupon is the same for all connections, regardless of degree. To see that the distribution in degree is the same for contacts as for the index case, regardless

of the degree of the respective index case, consider that the person referring a contact to participate to the research in conduct has $|(m-1)|$ options; the person whom previously participated in the research is excluded after all. The degree of the index case is assumed to be $|m|$. Then, a person in the list of contacts has a chance which scales with the number of coupons to be distributed divided by $|(m-1)|$ to be selected at random by the respective participant. In other words, the chance to receive a coupon at random scales with:

$$(m-1) \times \frac{\text{number of coupons}}{m-1} = \text{number of coupons}$$

And so, we conclude that the degree distribution of the person receiving a coupon does not differ from the distribution of the person handing out a coupon (given that they are from the same groups).

Of course, the exception on this occurs when there are more coupons to be distributed than contacts a person has. We do not explore this case any further.

2.2 Seeds and Waves

A seed is the first reachable participant to be included in the research. This seed might very well be of a certain, more accessible type. Through chain-referral sampling, one could also reach the group bearing the other trait. Getting the initial seed to supply information to the researcher and then having the researcher select a new participant will result in a new wave. In a way, one could state that the initial seeds are the $0th$ wave. Assuming only one coupon is being handed out, each new recruited participant makes for a new wave in the sample.

2.3 Coupons

Every participant in the research project receives a certain number of coupons which can be distributed among their respective contacts. Given a fixed number of participants, handing over a single coupon results in more waves than handing over several coupons. More coupons per participant might on the other hand allow gaining more information with a smaller chance of a chain 'ending'. Realistically, there is always a chance a chain could break, because a recruit with a coupon could choose not to proceed with the research, either by not participating or by refusing to contact another individual.

2.4 Degree

The degree of a subject is simply the number of contacts one has. We attempt to estimate the degree of all groups relevant to the research. Usually only two groups are being compared, and so, we limit ourselves to this particular case. A degree of a participant will be referred to as d_i.

The average degree of a group can be estimated as a mean of the degrees of every member. The mean degree of the group is as follows, with D_a being the mean degree of group A, d_i being the degree of person i, and $|A|$ being the total number of people in group A:

$$D_a = \frac{1}{|A|} \sum_{i=1}^{|A|} d_i$$

2.5 Prevalence

The prevalence is the proportion of the population being researched that has a certain trait of interest. This could be a personality type, it could be a disease (as common in epidemiology), but it could also for instance be an activity. However, it is always expressed as a proportion of a particular total population. Let us consider a random trait a, and its complementary trait b, for an undefined population. In this case, we can express the prevalence of a as follows:

$$P_a = \frac{\text{number of individuals with trait a}}{(\text{number of individuals with trait a}) + (\text{number of individuals with trait b})}$$

For estimating prevalence, we encounter several problems. The main problem, for which RDS is used as frequently as it is, is to obtain a sampling frame of populations with traits which make them difficult to contact. Moreover, it might not simply be possible to express the fraction of people of a sample with a certain trait in comparison to the entire sample due to homophily of contacts (more on this later). Basically, we can't assume that the sample retrieved is actually an entirely representative sample of the population. And so, we require to compensate for these biases in estimating the prevalence from a sample.

To estimate the prevalence, we make use of the reciprocity assumption [6]. As all relations are reciprocal, we have that the entire network is connected with ties between people in the population. This has the interesting implication that the number of ties from group A to group B is similar. So, take T_{ab}, the number of ties from group A to group B, and vice versa for T_{ba}. We get:

$$T_{ab} = T_{ba}$$

Obviously, this only holds on the population level. However, we can still make use of this approximation for an estimation. Moreover, the number of ties T_{ab} is equal to the mean degree of group A multiplied by the proportion of contacts of participants of group B and the prevalence of group A. A similar formula holds for T_{ba}. In short, we have:

$T_{ab} = P_a D_a p_{ab}$ and $T_{ba} = P_b D_b p_{ba}$.

With this we get the following equivalence:

$$P_a D_a p_{ab} = P_b D_b p_{ba} \tag{1}$$

Now, using $P_a + P_b = 1$, we can derive a formula for both P_a and P_b:

$$\begin{aligned} P_a &= \frac{D_b p_{ba}}{D_a p_{ab} + D_b p_{ba}} \\ P_b &= \frac{D_a p_{ab}}{D_a p_{ab} + D_b p_{ba}} \end{aligned} \tag{2}$$

2.6 Homophily

Participants with a certain trait could have a relative preference in contacts for people with the same trait (or opposite trait). Contacts might not be randomly distributed. This preference is commonly referred to as 'homophily of contacts'. For why this occurs, there are many possibilities.

Although there is agreement on this much, there are various formal definitions of homophily. One such definition, given by David Easley and Jon Kleinberg ([4]), compares the difference of cross-group connection likelihood with a null hypothesis of there being no significant preference. In short, this means that when we have a group A and a group B with respective proportions p and q, we infer that the null hypothesis would imply that the probability of a cross-group relation is $2pq$. This is reasonably intuitive, although possibly not entirely practical.

We intend to use the definition given by Douglas D. Heckathorn ([2]). In this approach, it is assumed that homophily implies a fixed proportion of contacts being of the same type. The other non-fixed contacts are randomly distributed over the other groups.

Assume that p_{aa} is the proportion of contacts of a member of group A that is of type A. Or rather, the chance thereof, as not all members of the group will have this same proportion; the chance of a contact in a certain group (first index) being of a certain type (second index). The chance of being selected as a person with 'trait a' depends on the degree of the respective groups as well. A group with a higher degree, is more likely to have the respective person as a contact than a group with a smaller degree, assuming equal prevalences. Of course, the prevalence is also relevant. All things considered, then, we get the following formula:

$$p_{aa} = H_a + (1 - H_a)\frac{P_a D_a}{P_a D_a + PbD_b}$$

In which H_a is the homophily of group A, P_a is the prevalence of 'trait a', and D_a is the (average) degree of group A.

We shall see that with this it follows that there is only one value for homophily which is equal for groups A and B, and therefore $H_a = H_b := H$. This follows from the result we get for the homophily of group A (which is the same as when the calculation is done for group B). We shall already refer to the homophily parameter as H. To this end, we shall also use that $p_{aa} = 1 - p_{ab}$ and $\frac{P_a D_a}{P_a D_a + P_b D_b} = (1 - \frac{P_b D_b}{P_a D_a + P_b D_b})$, latter as follows from algebra, by multiplying the 1 by the denominator of the other term.

We can now express the homophily in terms of the other quantities. We refer to the Appendix (23) for the calculation. Here we will only present the result, also see (24).

$$H = 1 - \frac{p_{ab}(P_a D_a + P_b D_b)}{P_b D_b}$$

And after filling in the prevalences P_a and P_b as derived in (2), we get (24):

$$H = 1 - p_{ba} - p_{ab}$$

We see from this result (24) that with random mixing, the homophily would be 0, as then $p_{ab} + p_{ba} = 1$. The maximum value is 1 and the minimum value is -1, as is evident from the formula.

3 Estimators

3.1 Degree

In order to estimate the degree, we have to take into account the fact that our sample is biased towards higher degree values. So a mean reported degree of the sample will not suffice. We will have to correct for this.

The issue here is that the degree of people in the sample are likely to be higher than that of the general population they were drawn from ([1]). A higher number of contacts increases the chance of being recruited after all, as there are more people who could refer to a person with a high degree. So, members with a high degree will then be oversampled. We can, however, weigh respondents inversely to their degree to counteract this bias.

For two individuals of which one has twice as many contacts as the other individual, the chance of being recruited for the person with more contacts is also twice as high [6]. By inversely correcting to their degree we can counteract this for the estimator. We suggest the following estimator, with N_A being the number of participants in group A, and $\tilde{d}_a$ being the estimated mean degree:

$$\tilde{d}_a = \frac{\sum_{i=1}^{N_A} \frac{1}{d_i} d_i}{\sum_{i=1}^{N_A} \frac{1}{d_i}} = \frac{N_A}{\sum_{i=1}^{N_A} \frac{1}{d_i}} \tag{3}$$

Now, to calculate the variance of the mean degree, we introduce $\tilde{\mu}_a = \frac{1}{\tilde{d}_a}$, as the harmonic mean is equivalent to the arithmetic mean of the reciprocals.

The variance of $\frac{1}{d_i}$ from the sample is as follows:

$$\sigma_A^2 = \frac{1}{N_A} \sum_{i=1}^{N_A} \left(\frac{1}{d_i} - \tilde{\mu}_a \right)^2 \tag{4}$$

We now wish to calculate the variance of the mean degree:

$$\begin{aligned}
Var(\tilde{d_a}) &= Var\left(\frac{N_A}{\sum_{i=1}^{N_A}\frac{1}{d_i}}\right) \\
&= Var\left(\frac{1}{\frac{1}{N_A}\sum_{i=1}^{N_A}\frac{1}{d_i}}\right) \\
&= Var\left(\frac{1}{\tilde{\mu}_a}\right)
\end{aligned}$$

(We expect $\tilde{\mu}_a$ to be close to μ_a. Therefore, we write $\tilde{\mu}_a = \mu_a + \epsilon_{\mu_a}$).

$$\begin{aligned}
&= Var\left(\frac{1}{\mu_a + \epsilon_{\mu_a}}\right) \\
&= Var\left(\frac{1}{\mu_a(1+\frac{\epsilon_{\mu_a}}{\mu_a})}\right) + \mathcal{O}((\frac{\epsilon_{\mu_a}}{\mu_a})^2) \\
&\approx \frac{1}{\mu_a^2} Var\left(1 - \frac{\epsilon_{\mu_a}}{\mu_a}\right) \\
&= \frac{1}{\mu_a^2} Var\left(\frac{\epsilon_{\mu_a}}{\mu_a}\right) \\
&= \frac{1}{\mu_a^4} Var\left(\epsilon_{\mu_a}\right) \\
&= \frac{1}{\mu_a^4} Var\left(\epsilon_{\mu_a} + \mu_a\right) \\
&= \frac{1}{\mu_a^4} Var\left(\frac{1}{N_A}\sum_{i=1}^{N_A}\frac{1}{d_i}\right) \\
&= \frac{1}{\mu_a^4 N_A^2}\sum_{i=1}^{N_A} Var\left(\frac{1}{d_i}\right) \\
&= \frac{1}{\mu_a^4 N_A^2} N_A \sigma_A^2 \\
&= \frac{1}{\mu_a^4 N_A}\sigma_A^2
\end{aligned} \tag{5}$$

To get an estimator for the variance of the degree of the population, we use the following expression:

$$Var(\tilde{d_a}) = \frac{1}{N_A}\frac{\sigma_A^2}{\tilde{\mu}_a^4} \tag{6}$$

Now, with a large enough sample (so that N_A becomes large enough), we get the following normal distribution:

$$\tilde{d_a} \sim N\left(D_a, \frac{1}{N_A}\frac{\sigma_A^2}{\tilde{\mu}_a^4}\right) \tag{7}$$

3.2 Prevalence

In order to estimate the prevalence of a group of a certain type / with a certain characteristic, we make the distinction between several situations. First of, we want to know the estimators in the case of uniform sampling. We wish to explore the estimators for non-uniform sampling as well, with and without perfect information retrieved from participants with coupons about their contacts.

3.2.1 Uniform Sampling through coupons

We start off by estimating the prevalence in the case of uniform sampling through coupons. We do not use the information retrieved from index cases about contacts.

First of, to estimate the chance of a contact being of a certain type within a certain group, also for now referred to as selection or contact proportions, we require the use of the recruitment counts, which are strictly speaking also estimators, but we are not interested in them. We denote these by k_{aa}, k_{ab}, k_{ba} and k_{bb}. Now, we estimate the type contact or selection proportion of contacts of type B selected by type A as:

$$\tilde{p}_{ab} = \frac{k_{ab}}{k_{aa} + k_{ab}}$$

Given sufficiently large values of k_{aa} and k_{ab}, the selection proportion is normally distributed with mean p_{aa} (the true selection proportion) and corresponding variance, i.e.:

$$\tilde{p}_{aa} \sim N(p_{aa}, \frac{p_{aa}(1-p_{aa})}{k_{aa} + k_{ab}})$$

Conceptually, we can see that homophily and mean degree influence the retrieved recruits in the sample. We already derived a relation for the prevalence based on this theory, see (2). Of course, it is assumed that the sample size is big enough. Now, through an estimation of the degree and the selection proportions, we can derive one for the prevalence:

$$\tilde{p}_a = \frac{\tilde{d}_b \tilde{p}_{ba}}{\tilde{d}_a \tilde{p}_{ab} + \tilde{d}_b \tilde{p}_{ba}} \tag{8}$$

Now that we have an estimation for the prevalence, we also wish to know the variance of the error of the estimator. We consider the following function:

$$f(\tilde{d}_a, \tilde{d}_b, \tilde{p}_{ab}, \tilde{p}_{ba}) = \frac{\tilde{d}_b \tilde{p}_{ba}}{\tilde{d}_a \tilde{p}_{ab} + \tilde{d}_b \tilde{p}_{ba}}$$

To calculate the variance of $\tilde{p}_a$, we take a point $\theta = (\mu_1, \mu_2, \mu_3, \mu_4)$ as estimator, and add an error term to all these variables. Essentially what we do is working out the following Taylor expansion:

$$f(\mu_1 + \epsilon_1, \mu_2 + \epsilon_2, \mu_3 + \epsilon_3, \mu_4 + \epsilon_4) = f(\mu_1, \mu_2, \mu_3, \mu_4) + \sum_{i=1}^{4} \frac{\partial f}{\partial x_i}(\mu_1, \mu_2, \mu_3, \mu_4) \times \epsilon_i + \mathcal{O}(\epsilon_i \epsilon_j)$$

As a result, we get the following variance (see the Appendix for the full calculation):

$$Var(\tilde{p}_a) = (-\frac{\tilde{p}_a \tilde{p}_b}{\tilde{d}_a})^2 \times Var(\epsilon_1) + (\frac{\tilde{p}_a \tilde{p}_b}{\tilde{d}_b})^2 \times Var(\epsilon_2) + (-\frac{\tilde{p}_a \tilde{p}_b}{\tilde{p}_{ab}})^2 \times Var(\epsilon_3) + (\frac{\tilde{p}_a \tilde{p}_b}{\tilde{p}_{ab}})^2 \times Var(\epsilon_4) \tag{9}$$

We recognise the terms ϵ_1 and ϵ_2 as respectively the errors of the mean degree of group A and B, while ϵ_3 and ϵ_4 are the errors of $\tilde{p}_{ab}$ and $\tilde{p}_{ba}$.

The variance of the mean degree gives the following results:

$$Var(\epsilon_1) = \frac{1}{\mu_a^4} \frac{\sigma_A^2}{N_A}$$

$$Var(\epsilon_2) = \frac{1}{\mu_b^4} \frac{\sigma_B^2}{N_B}$$

With σ_A and σ_B being the standard deviation of the reciprocal of sample mean degree of group A and B, and N_A and N_B being respectively the number of group A and B recruits, see 4.

To estimate the variance of ϵ_3 and ϵ_4, we use the distributions of $\tilde{p}_{ab}$ and $\tilde{p}_{ba}$:

$$\tilde{p}_{ab} \sim N\left(\frac{k_{ab}}{k_{aa}+k_{ab}}, \frac{\tilde{p}_{aa}\tilde{p}_{ab}}{k_{aa}+k_{ab}}\right)$$

The error of the selection proportion (or better yet: chance of a contact being of a certain type within a certain group) is normally distributed with the same variance as $\tilde{p}_{ab}$:

$$Var(\epsilon_3) = \frac{\tilde{p}_{aa}\tilde{p}_{ab}}{k_{aa}+k_{ab}}$$

$$Var(\epsilon_4) = \frac{\tilde{p}_{ab}\tilde{p}_{bb}}{k_{ab}+k_{bb}}$$

3.2.2 Using contact information from index cases

Although it is a common practice in RDS to only make use of information supplied by the participant receiving a coupon, there is also the possibility to gain information about all contacts of a participant at once. After all, an index case is supposed to know all of its contacts, and the participants are asked about their degree. This would drastically increase the amount of information gained per coupon, scaling with the mean degree, if given information is entirely accurate. However, if this information can not be supplied with absolute certainty, an estimation by the index case about the contacts has to be made. From the sample gained, we can ascertain whether the information about traits are correct for the participants which received coupons, and compare it with the estimation of the respective contact which gave the coupon and made an estimation. Now, if this estimation is no better than guessing, we do not expect to receive any valuable information from using all information given about contacts. In fact, this information will then essentially be useless. Given the correlation calculated from the sample, we would like to know at what point the data is too contaminated with uncertainty for it to be of value. We will discuss that more in 3.2.6.

For the non-uniform cases, it is essential to know the contact information of every index case. We are not, however, discussing non-uniform sampling here. We refer to formula 9 for the relevant equation. Our information about types (which we notified as recruitment counts before in the uniform sampling case) will more or less scale with the (size biased) degree of participants.

As for the estimators, take the recruitment counts to be the counts of a certain type within a certain group instead. Of course, these counts will be far higher than in the uniform case we treated before, simply for the reason that we have to retrieve information about all contacts to select new participants.

Now, to derive a variance to account for the increased amount of information we retrieve, we refer to 9. In this equation, we witness a sum of four terms. With a set degree, we attain that the first two terms with $Var(\epsilon_1)$ and $Var(\epsilon_2)$ are 0, as the values of σ_A and σ_B become 0 with a set degree. We then see that the other two terms dominate, but the increase of information gained from using the information about contacts, will result in higher numerators: $k_{aa}+k_{ab}$ and $k_{ab}+k_{bb}$. If we take this increase of information, but continue sampling uniformly, the variance terms $Var(\epsilon_3)$ and $Var(\epsilon_4)$ become smaller by (the size unbiased degree - 1 - #coupons); in this case the set degree of $10-1-1$ that we considered. The 1 stands for the person the coupon was received from, and we would already possess the information about the people who are about to receive their coupons as well. The scaling factor between merely coupon information and full contact information is then as follows:

$$\frac{\text{size biased degree} - 1}{\#\text{ coupons}} \tag{10}$$

With a variable degree, we will get the situation that the first two variance terms will dominate, as they will not be equal to 0. Though the variance of the prevalence reduces by degree, the variance terms involving the recruitment counts will be so small eventually if the degree is sufficiently large, that an even larger degree is not going to make a significant difference.

We will cover an example to clarify. Assume the degree is constant. In this case, we see from 9, that the information of all waves except the last and the seed ($0th$ wave) will result in a variance of the prevalence which is smaller by the term as given in Equation 10. We will also treat a variable degree, a variance of 2, equal prevalences, and a sample size of 50 for each group. Note that the parameters we have chosen are very

artificial. The Python code for the example can be found in the Appendix at 5.3. The parameters for the example can also; we have chosen very generic parameters for the sake of giving an example of approximately how to do a comparison between full contact information and no contact information.

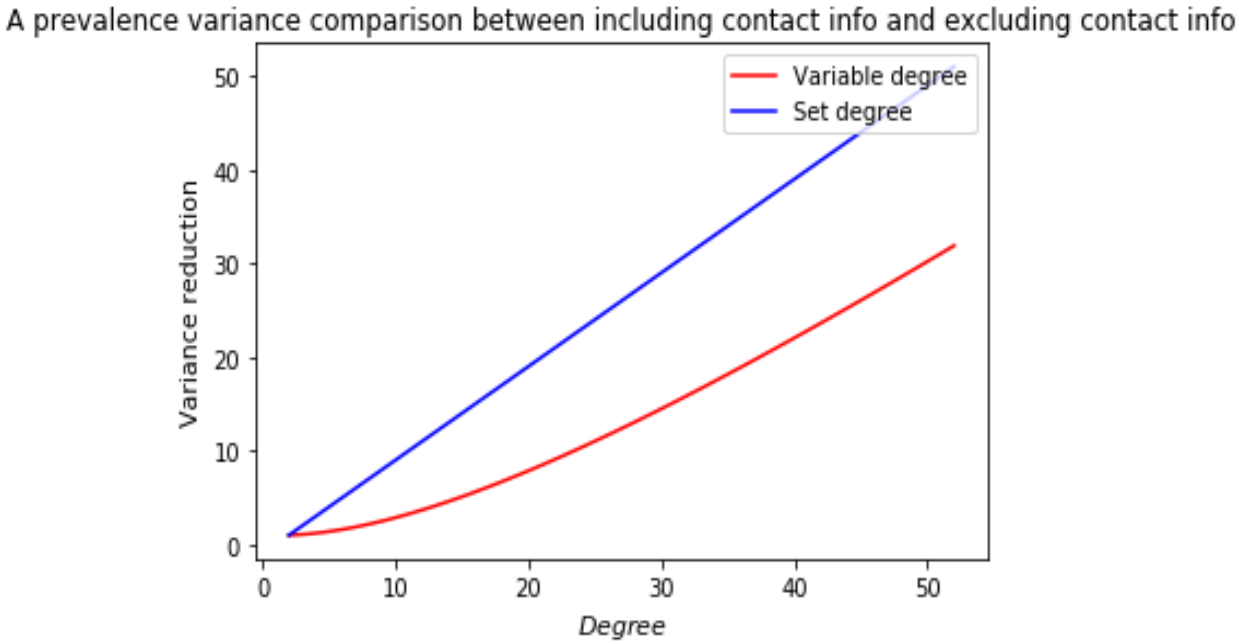

Figure 1: A generic example of a prevalence variance comparison between full contact information and exclusively coupon information. The functions plot a ratio between full contact info and no contact info. The red line (variable degree) has set mu and sigma parameters, indicative of a variable degree. The blue line only involves the epsilon3 and epsilon4 terms in the expression for the variance of the prevalence; these are variances of the probability of a contact being of a certain type within a certain group.

Of course, as the μ_a, μ_b, σ_A and σ_B parameters are so artificially chosen, this is not a very realistic example, but it serves to get the idea of doing a variance comparison across.

3.2.3 Non-Uniform Sampling: Perfect Information

3.2.4 Introduction

As the case of non-uniform sampling is comparable to the case above with uniform sampling, except for having a greater sample size and different sampling, we can derive the relevant results using the previous section.

Our information about types (which we notified as recruitment counts before) will scale with the (size biased) degree of participants. However, as we also sample non-uniformly between the two groups, we can have a similar number of participants from one group with a coupon as to the other quite rapidly, with little uncertainty. That is, in the non-uniform case, we can establish to hand a coupon to a group with a probability of a half (given that both groups are among the contacts of every index case). This works well mainly when the prevalences are actually about equal. It bypasses the problems homophily presents. This would mean that even with a small sample, we reach an even amount of both groups with a reasonably small error, conceptually speaking. See Figure 2.

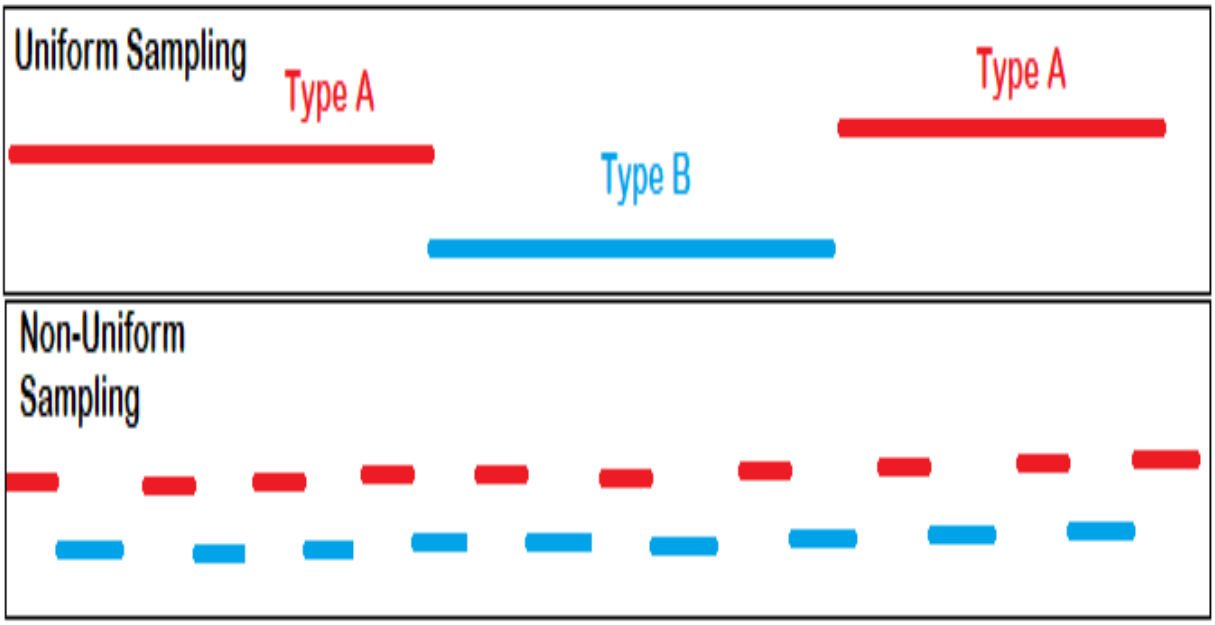

Figure 2: Due to sampling non-uniformly, the chains of similar types with a coupon in the sample get much smaller. This ultimately results in less variance in the total number of participants recruited from both types.

There are issues with this sampling method. Assume that the only relevant group is A (that is, B is essentially non-existent or has a prevalence of less than a percent), and we estimate prevalences of A and B to be about equal. In this case, one would expect to retrieve a lot of data about this insignificant group B, and non-uniform sampling will be essentially twice as bad as uniform sampling, as an upperbound. However, if the degree of the participants is finite and sufficiently low (say, about 10), it may well be that participants of type B do not even appear as contacts of any participant. So, in reality this issue might not be as staggering as an upperbound to the reduction of the sample size from the sampling method by 2. The issue with this, is that all the information gained about the smaller group will hardly reduce the variance of the prevalence, see (9).

To elaborate, imagine there is homophily, and we begin sampling from group A. Now, in the case of uniform sampling, we could get a chain of participants from group A, before we jump to a chain of group B. It is even possible, that we never get any participants from group B. Essentially, the issue is that we cannot reach the desired proper mixing with a small relative error (or variance rather). This can potentially result in a much higher variance of the prevalence from the retrieved sample than anticipated.

What we desire is to compare the advantage of non-uniform sampling compared to uniform sampling when it comes to attaining data from a sample. Of course, we have to take into account, that for non-uniform sampling, it is necessary to retrieve data of all contacts of the index cases, because we need that to select our new recruits.

Now, as for the variance in terms of attained types of the index cases, we point out that we select them with a set probability, and therefore always get a set distribution with a set binomial variance, given of course that the degree and prevalences are sufficiently large. The number of participants in group A with non-uniform sampling, which is indicated by N_A, has a variance, assuming a probability of a half to switch group, of approximately (binomially distributed):

$$Var(\tilde{N}_A) = 1/2 \times (1 - 1/2) \times \#\text{waves} = \#\text{waves}/4$$

3.2.5 In practice

The probability that an index case of type A includes a person of type B is equal to:

$$0\times P(\text{Type A no contact with type B})+\frac{1}{2}P(\text{Type A contact with A and B})+1\times P(\text{Type A no contact with A})$$

Assuming that individual i has a degree d_i, and the chance of an individual of type A to have contact with an individual of type B is p_{ab}, then we get the following probability, assuming that the index case can not return the coupon back to whoever gave it:

$$0 \times (1-p_{ab})^{d_i-1} + \frac{1}{2} \times (1-(1-p_{ab})^{d_i-1} - (p_{ab}^{d_i-1})) + 1 \times p_{ab}^{d_i-1}$$

In general, the probability of a connection from A to B could be given as:

$$\sum_{i=2}^{\infty} \frac{iP_i}{\sum_{j=2}^{\infty} jP_j} \left[\frac{1}{2}(1-(1-p_{ab})^{i-1} - p_{ab}^{i-1}) + p_{ab}^{i-1} \right] \tag{11}$$

In which P_i is a probability of retrieving a certain degree. By setting the degree for all participants on 10, we can get an intuitive idea of what to retrieve; it will be merely an exercise in bookkeeping to tackle more possibilities in practice. Based on the parameters (proportions or probabilities of contacts being of a certain type within a certain type), we can then compute the matrix with transition probabilities per wave.

With a set degree, we can factor out these estimators. We have to note, however, that the increase of information gained from contacts does not decrease the estimator of the degree. However, the sampling method does alter the number of participants retrieved in both groups, and therefore can affect the variance of the prevalence.

Also with a set degree, we are left with $\frac{\tilde{p}_{ba}}{\tilde{p}_{ab}+\tilde{p}_{ba}}$. The increase of information, leads to an expression in which we get a Bernoulli trial per contact within the sample of type A of their contacts, to verify whether they are of type B or not.

$$\frac{\sum_{i=1}^{\tilde{N}_b} \sum_{j=1}^{D_i-1} Ber(p_{ba})}{\sum_{i=1}^{\tilde{N}_b} \sum_{j=1}^{D_i-1} 1} \tag{12}$$

Comparing non-uniform sampling and uniform sampling, will involve the probability distribution of retrieving a number of contacts of a certain type, and integrating or summing the variance of the prevalence over all possibilities and probability values.

These concepts can be used for simulations.

3.2.6 Non-Uniform Sampling: Imperfect Information

This is the most complicated and interesting case. We got a situation in which we attempt to estimate the prevalence of a type through information about types and reports of index cases, however the reports are not guaranteed to be 100% correct. In essence, the data we possess is the types of the participants who have been handed out a coupon, and the reports of every index cases about their respective contacts.

For simplicity's sake, let us label these C_{AA} for the number of participants with a coupon, which happen to be of type A, and were estimated correctly to be of type A by a contact. We label C_{AB} for the participants with a coupon which were of type B and were reported to be of type A, and accordingly, we label participants of type A which were reported to be of type B as C_{BA}, and participants of type B which were reported as B as C_{BB}. Moreover, contacts which have not received a coupon, but were reported to be of type A will be labeled as R_A, and contacts which were reported to be of type B as R_B.

Through this information, we can estimate a sensitivity parameter S_A and S_B, which stand respectively for the chance of estimating a contact to be of type A given that the respective contact is of type A, and the chance of estimating a contact to be of type B, given that the person is of type B. In other words:

$$S_A = P(\text{Person reported to be of type A}|\text{Type A})$$

$$S_B = P(\text{Person reported to be of type B}|\text{Type B})$$

Then we have the following approximations, which we will try to use when linearising the log-likelihood function we will compose.

$$S_A \approx \frac{C_{AA}}{C_{AA} + C_{BA}} \tag{13}$$

$$S_B \approx \frac{C_{BB}}{C_{BB} + C_{AB}} \tag{14}$$

The two other parameters we have, for participants of type A, are $\tilde{p}_{aa}$ and $\tilde{p}_{ab} = 1 - \tilde{p}_{aa}$. We have to estimate these using the data we have. Then we can compute the prevalence again as we did before, using the theory from (2). We are not interested in the sensitivity parameters, but we have to use the approximations to get a solvable system of equations with Mathematica. Because if we were to attempt taking the derivate of the LogLikelihood function to every parameter and setting it equal to 0, we would not get a solvable system.

Now, as for estimating the proportions from the data we have and the sensitivities, we use a log-likelihood approach. The likelihood function will be as follows:

$$\begin{aligned} L(\tilde{p}_{AA}, S_A, S_B) = \Big(\tilde{p}_{AA}S_A\Big)^{C_{AA}} \times \\ \Big((1-\tilde{p}_{AA})(1-S_B)\Big)^{C_{AB}} \times \Big(\tilde{p}_{AA}(1-S_A)\Big)^{C_{BA}} \times \\ \Big((1-\tilde{p}_{AA})S_B\Big)^{C_{BB}} \times \\ \Big(\tilde{p}_{AA}S_A + (1-\tilde{p}_{AA})(1-S_B)\Big)^{R_A} \times \Big((1-\tilde{p}_{AA})S_B + \tilde{p}_{AA}(1-S_A)\Big)^{R_B} \end{aligned} \tag{15}$$

This naturally results in the following log-likelihood function:

$$\begin{aligned} \log L(\tilde{p}_{AA}, S_A, S_B) = C_{AA}\Big(\log(\tilde{p}_{AA}) \log(S_A)\Big) + \\ C_{AB}\Big(\log(1-\tilde{p}_{AA}) + \log(1-S_B)\Big) + C_{BA}\Big(\log(\tilde{p}_{AA}) \log(1-S_A)\Big) + \\ C_{BB}\Big(\log(1-\tilde{p}_{AA}) + \log(S_B)\Big) + \\ R_A \log\Big(\tilde{p}_{AA})S_A + (1-\tilde{p}_{AA})(1-S_B)\Big) + R_B \log\Big((1-\tilde{p}_{AA})S_B + \tilde{p}_{AA}(1-S_A)\Big) \end{aligned} \tag{16}$$

Now, additionally, using the following three approximations of the coupon information, we can solve with Mathematica to retrieve an estimate of the proportion of people of type A in the sample of type A. Note that Nc_A indicates the number of coupons given by type A individuals, so that equals $C_{AA} + C_{BA} + C_{AB} + C_{BB}$. See the Appendix for the full calculations 5.4.

$$C_{AA} = \tilde{p}_{aa}S_A Nc_A$$
$$C_{AB} = (1-\tilde{p}_{aa})(1-S_B)Nc_A$$

$$C_{BB} = (1 - \tilde{p}_{aa}) S_B N c_A$$

We retrieve the aforementioned 13 and 14, but also:

$$\tilde{p}_{aa} = \frac{C_{AA} + C_{BA}}{N c_A}$$

Now, linearising the parameters of the LogLikelihood Function (16) by adding an error term ϵ_A to S_A, ϵ_B to S_B, and an error term ϵ_p to $\tilde{p}_{aa}$, and then calculating consecutive series to the first order, after which we simplify the result, we retrieve an expression which can be found in the Appendix at 5.4. The result of the series expansion after inserting the approximations and derivating to $\tilde{p}_{aa}$ is as follows:

Figure 3: The simplified series for the solution of $\tilde{p}_{aa}$.

We take a different approach now. We once again refer to the LogLikelihood Function 16. This time we will manually derive the results, to retrieve expression for the variance terms of the sensitivity and the probability of type A within group A. However, this time, we will use only one sensitivity parameter: S.

So, once again, we have: 16. We will name this function l.

From this, we retrieve the following two derivates:

$$\frac{\partial l}{\partial S} = \frac{C_{AA} + C_{BB}}{S} - \frac{C_{AB} + C_{BA}}{1 - S} + \frac{R_A(2\tilde{p}_{aa} - 1)}{\tilde{p}_{aa} \times S + (1 - \tilde{p}_{aa})(1 - S)} + \frac{R_B(1 - 2\tilde{p}_{aa})}{(1 - \tilde{p}_{aa}) \times S + \tilde{p}_{aa} \times (1 - S)} \quad (17)$$

$$\frac{\partial l}{\partial \tilde{p}_{aa}} = \frac{C_{AA} + C_{BA}}{\tilde{p}_{aa}} - \frac{C_{AB} + C_{BB}}{1 - \tilde{p}_{aa}} + \frac{R_A(2S - 1)}{\tilde{p}_{aa} \times S + (1 - \tilde{p}_{aa})(1 - S)} + \frac{R_B(1 - 2S)}{(1 - \tilde{p}_{aa}) \times S + \tilde{p}_{aa} \times (1 - S)} \quad (18)$$

Now, for both derivatives, we use the following series approximation:

$$S = S_0 + \epsilon_s$$

$$p_{aa} = p_{aa}^0 + \epsilon_p$$

And using the following Taylor series:

$$\frac{1}{S} = \frac{1}{S_0 + \epsilon_s} = \frac{1}{S_0}\left(\frac{1}{1 + \frac{\epsilon_s}{S_0}}\right) = \frac{1}{S_0}\left(1 - \frac{\epsilon_s}{S_0}\right)$$

We can express the first two terms in 17 as:

$$\frac{C_{AA} + C_{BB}}{S_0}\left(1 - \frac{\epsilon_s}{S_0}\right) - \frac{C_{AB} + C_{BA}}{1 - S_0}\left(1 - \frac{\epsilon_s}{1 - S_0}\right)$$

And in 18 as:

$$\frac{C_{AA} + C_{BA}}{p^0_{aa}}\left(1 - \frac{\epsilon_p}{p^0_{aa}}\right) - \frac{C_{AB} + C_{BB}}{1 - p^0_{aa}}\left(1 - \frac{\epsilon_p}{1 - p^0_{aa}}\right)$$

Now, for the third term, after replacing the parameters by the approximations, and expanding the denominator, we retrieve the following expression:

$$\frac{R_A(2p^0_{aa} + 2\epsilon_p - 1)}{p^0_{aa}S_0 + (1 - p^0_{aa})(1 - S_0) + \epsilon_p(2S_0 - 1) + \epsilon_s(2p^0_{aa} - 1) + \mathcal{O}(\epsilon_s\epsilon_p)}$$

We label the constant term as K:

$$K = (p^0_{aa}S_0 + (1 - p^0_{aa})(1 - S_0))$$

And with this constant term, we retrieve:

$$\frac{R_A(2p^0_{aa} - 1 + 2\epsilon_p)}{K(1 + \frac{\epsilon_p(2S_0 - 1)}{K} + \frac{\epsilon_s(2p^0_{aa} - 1)}{K}) + \mathcal{O}(\epsilon_s\epsilon_p)}$$

Now we apply a Taylor series:

$$\frac{R_A(2p^0_{aa} - 1 + 2\epsilon_p)}{K}\left(K(1 - \frac{\epsilon_p(2S_0 - 1)}{K} - \frac{\epsilon_s(2p^0_{aa} - 1)}{K}) + \mathcal{O}(\epsilon_s\epsilon_p)\right)$$

And with that, we can find the linearised expression:

$$\frac{R_A(2p^0_{aa} - 1)}{K} + \epsilon_p\left(\frac{2R_A}{K} - \frac{2S_0 - 1}{K}\frac{R_A(2p^0_{aa})}{K}\right) + \epsilon_s\left(-\frac{2p^0_{aa} - 1}{K}\frac{R_A(2p^0_{aa} - 1)}{K}\right) + \mathcal{O}(\epsilon^2)$$

Additionally, through similar methods, and by defining the constant:

$$K' = (S_0 + p^0_{aa} - 2S_0p^0_{aa})$$

We get:

$$\frac{R_B(1 - 2p^0_{aa})}{K'} + \epsilon_p\left(\frac{-2R_B}{K'} - \frac{(1 - 2S_0)R_B(1 - 2p^0_{aa})}{K'^2}\right) + \epsilon_s\left(\frac{1 - 2p^0_{aa}}{K'}R_B\frac{1 - 2p^0_{aa}}{K'}\right)$$

Now, combining all these terms, we get the overall linearisation of 17:

$$A + B \times \epsilon_s + C \times \epsilon_p = 0 \tag{19}$$

With:

$$A = \frac{R_B(1 - 2p^0_{aa})}{K'} + \frac{R_A(2p^0_{aa} - 1)}{K} + \frac{C_{AA} + C_{BB}}{S_0} - \frac{C_{AB} + C_{BA}}{1 - S_0}$$

$$B = \frac{R_B(1 - 2p^0_{aa})^2}{K'^2} + \frac{R_A(2p^0_{aa} - 1)^2}{K^2} - \frac{C_{AA} + C_{BB}}{S_0^2} - \frac{C_{AB} + C_{BA}}{(1 - S_0)^2}$$

$$C = \left(\frac{-2R_B}{K'} - \frac{R_B(1 - 2S_0)(1 - 2p^0_{aa})}{K'^2}\right) + \left(\frac{2R_A}{K} - \frac{2S_0 - 1}{K}\frac{R_A(2p^0_{aa} - 1)}{K}\right)$$

We require that we set this equation 19, which is the linearised version of 17, equal to 0.

Similarly, we can retrieve a linearised version of 18 through similar methods. We will not go through the steps.

Expressing the third term as:

$$\frac{R_A(2S_0-1)}{K}+\epsilon_s\left(\frac{-2R_A}{K}-\frac{2p^0_{aa}-1}{K}\frac{R_A(2S_0-1)}{K}\right)+\epsilon_p\left(-\frac{(2S_0-1)}{K}\frac{R_A(2S_0-1)}{K}\right)+\mathcal{O}(\epsilon^2)$$

And the fourth term as:

$$\frac{R_B(1-2S_0)}{K'}+\epsilon_s\left(\frac{1-2S_0}{K'}R_B\frac{1-2p^0_{aa}}{K'}\right)+\epsilon_p\left(\frac{-2R_B}{K'}-(1-2S_0)^2\frac{R_B}{K'^2}\right)+\mathcal{O}(\epsilon^2)$$

And with that, we get the linearised expression of 18:

$$D+E\times\epsilon_s+F\times\epsilon_p=0 \tag{20}$$

With:

$$\begin{aligned}
D&=\frac{C_{AA}+C_{BA}}{p^0_{aa}}-\frac{C_{AB}+C_{BB}}{1-p^0_{aa}}+\frac{R_A(2S_0-1)}{K}+\frac{R_B(1-2S_0)}{K'}\\
E&=\frac{-2R_A}{K}-\frac{2p^0_{aa}-1}{K}\frac{R_A(2S_0-1)}{K}+\frac{1-2S_0}{K'}R_B\frac{1-2p^0_{aa}}{K'}\\
F&=-\frac{2S_0-1}{K}\frac{R_A(2S_0-1)}{K}+\left(\frac{-2R_B}{K'}-\frac{(1-2S_0)^2R_B}{K'^2}\right)
\end{aligned}$$

Now we can combine 19 and 20 to solve for the epsilon values, or also the variance terms of the sensitivity and contact type probability parameters. We then find the following solutions:

$$\epsilon_s=-\frac{AF-DC}{FB-EC} \tag{21}$$

$$\epsilon_p=-\frac{AE-DB}{CE-BF} \tag{22}$$

4 Discussion

Although we formulated several methods and presented the formulas with which a researcher could investigate the benefits of non-uniform sampling, we can not conclusively prove its merits. However, given suitable prerequisites, the decreased variance in type sampling can notably reduce the variance of the prevalence estimator. Specific examples or estimates of situations will have to be checked accordingly, analytically or through simulations. If a researcher aims for a certain maximum variance, it can be accomplished with a higher certainty with non-uniform sampling than with uniform sampling, as with the latter we get the issue that there's always a legitimate chance of undersampling one of the groups.

Potentially interesting to investigate could be the situation of several groups existing in a certain population, with several of the groups being hard-to-reach. Essentially, this will be a more elaborate case of the same thing, with once again an equal number of ties between all groups. The prevalence can then be expressed in several ways, as we can take the degree of different groups, and the estimator which results in the lowest variance would be optimal.

Given reasonable conditions, non-uniform sampling is a potent method to retrieve results of estimators such as prevalence with more precision than can be retrieved with uniform sampling. Although it has its issues, most crucially with the way a probability of sampling between groups has to be established beforehand, it has the potential to improve upon the current sampling methods applied with RDS.

A further suggestion of interest could be to investigate whether the sampling probability between groups can be adjusted by interim variance calculations of the sample at every wave. This way, the variance of the prevalence can reduced ongoing by decreasing the biggest variance term at every wave.

5 Appendix

5.1 Homophily

$$
\begin{aligned}
p_{aa} &= H + (1-H)\frac{P_aD_a}{P_aD_a+P_bD_b} \\
p_{aa} - \frac{P_aD_a}{P_aD_a+P_bD_b} &= H(1-\frac{P_aD_a}{P_aD_a+P_bD_b}) \\
H &= \frac{p_{aa} - \frac{P_aD_a}{P_aD_a+P_bD_b}}{(1-\frac{P_aD_a}{P_aD_a+P_bD_b})} \\
&= \frac{(1-p_{ab}) - (1-\frac{P_bD_b}{P_aD_a+P_bD_b})}{(\frac{P_bD_b}{P_aD_a+P_bD_b})} \\
&= \frac{-p_{ab} + \frac{P_bD_b}{P_aD_a+P_bD_b}}{(\frac{P_bD_b}{P_aD_a+P_bD_b})} \\
&= 1 - \frac{p_{ab}(P_aD_a+P_bD_b)}{P_bD_b}
\end{aligned}
\tag{23}
$$

We now fill in the prevalences P_a and P_b as derived in 2:

$$
\begin{aligned}
H &= 1 - \frac{p_{ab}(\frac{D_bp_{ba}}{D_ap_{ab}+D_bp_{ba}}D_a + \frac{D_ap_{ab}}{D_ap_{ab}+D_bp_{ba}}D_b)}{\frac{D_ap_{ab}}{D_ap_{ab}+D_bp_{ba}}D_b} \\
&= 1 - \frac{p_{ab}D_bp_{ba}D_a + p_{ab}D_ap_{ab}D_b}{D_aD_bp_{ab}} \\
&= 1 - p_{ba} - p_{ab}
\end{aligned}
\tag{24}
$$

5.2 Variance Prevalence

For the variance of the estimator of the prevalence $\tilde{p}_a$, we consider a function f of variables $\tilde{d}_a$, $\tilde{d}_b$, $\tilde{p}_{ab}$, and $\tilde{p}_{ba}$:

$$
f(\tilde{d}_a, \tilde{d}_b, \tilde{p}_{ab}, \tilde{p}_{ba}) = \frac{\tilde{d}_b\tilde{p}_{ba}}{\tilde{d}_a\tilde{p}_{ab} + \tilde{d}_b\tilde{p}_{ba}}
$$

We approximate the prevalence in point $\theta = (\mu_1, \mu_2, \mu_3, \mu_4)$. We use the notation that $\delta = \mu_1\mu_3 + \mu_2\mu_4$.

$$
f(\tilde{d}_a, \tilde{d}_b, \tilde{p}_{ab}, \tilde{p}_{ba}) = \frac{(\mu_2+\epsilon_2)(\mu_4+\epsilon_4)}{(\mu_1+\epsilon_1)(\mu_3+\epsilon_3) + (\mu_2+\epsilon_2)(\mu_4+\epsilon_4)}
$$

We use the notation that $\delta = \mu_1\mu_3 + \mu_2\mu_4$

$$
\begin{aligned}
&= \frac{1}{\delta} \times \frac{\mu_2\mu_4 + \mu_4\epsilon_2 + \epsilon_2\mu_4 + \mathcal{O}(\epsilon_2\epsilon_4)}{1 + \frac{\mu_3\epsilon_1}{\delta} + \frac{\mu_1\epsilon_3}{\delta} + \frac{\mu_4}{\delta} + \frac{\mu_2\epsilon_4}{\delta} + \mathcal{O}(\epsilon^3)} \\
&= \frac{1}{\delta} \times (\mu_2\mu_4 + \mu_2\epsilon_2 + \mu_2\epsilon_4 + \mathcal{O}(\epsilon_2\epsilon_4))(1 - \frac{\mu_3\epsilon_1}{\delta} - \frac{\mu_1\epsilon_3}{\delta} - \frac{\mu_4\epsilon_2}{\delta} - \frac{\mu_2\epsilon_4}{\delta} - \mathcal{O}(\epsilon^3)) \\
&\approx \frac{\mu_2\mu_4}{\delta} + \frac{\mu_4\epsilon_2}{\delta} + \frac{\mu_2\epsilon_4}{\delta} - \frac{\mu_2\mu_4}{\delta}(\frac{\mu_3\epsilon_1}{\delta} + \frac{\mu_1\epsilon_3}{\delta} + \frac{\mu_4\epsilon_2}{\delta} + \frac{\mu_2\epsilon_4}{\delta}) \\
&= \frac{\mu_2\mu_4}{\delta} + \epsilon_1(-\frac{\mu_3\mu_2\mu_4}{\delta^2}) + \epsilon_2\frac{\mu_4}{\delta}(1 - \frac{\mu_2\mu_4}{\delta}) + \epsilon_3(-\frac{\mu_1\mu_2\mu_4}{\delta^2}) + \epsilon_4\frac{\mu_2}{\delta}(1 - \frac{\mu_2\mu_4}{\delta})
\end{aligned}
$$

Assume that $c = \frac{\mu_2\mu_4}{\delta}$. We write the other terms before the epsilon values as constants c_i too.

$$
= c + \epsilon_1c_1 + \epsilon_2c_2 + \epsilon_3c_3 + \epsilon_4c_4 + \mathcal{O}(\epsilon^3)
$$

Now to take the variance:

$$Var(\tilde{p}_a) = Var(c + \epsilon_1 c_1 + \epsilon_2 c_2 + \epsilon_3 c_3 + \epsilon_4 c_4)$$

Because of independence:

$$Var(\tilde{p}_a) = c_1^2 Var(\epsilon_1) + c_2^2 Var(\epsilon_2) + c_3^2 Var(\epsilon_3) + c_4^2 Var(\epsilon_4)$$

As all c_i terms are known constants, only the variance of the ϵ_i's have to be found. Replacing all c_i's using the relation found in (2) gives:

$$Var(\tilde{p}_a) = (-\frac{\tilde{p}_a\tilde{p}_b}{\tilde{d}_a})^2 * Var(\epsilon_1) + (\frac{\tilde{p}_a\tilde{p}_b}{\tilde{d}_b})^2 * Var(\epsilon_2) + (-\frac{\tilde{p}_a\tilde{p}_b}{\tilde{p}_{ab}})^2 * Var(\epsilon_3) + (\frac{\tilde{p}_a\tilde{p}_b}{\tilde{p}_{ab}})^2 * Var(\epsilon_4)$$

5.3 Simulation for including full contact information

```
import matplotlib.pyplot as plt
import numpy as np

# only 1 coupon is being used

pa = 1/2
pb = 1/2

paa = 1/2
pbb = 1/2
pab = 1/2
pba = 1/2

Na = 50
Nb = 50

kaa = 25
kab = 25
kba = 25
kbb = 25

sigmaA = 0.01
sigmaB = 0.01

'''
Var1 = (1/mu^4 * sigmaA^2 / Na)
Var2 = (1/mu^4 * sigmaB^2 / Nb)
Var3 = paa*pbb / (kaa + kab)
Var4 = pab*pbb / (kab + kbb)
'''

'''
# Because sigma involves mu, and we take a set sigma value,
but do not want to blow up the actual variance,
we will work with a set value for mu too.

See the formula for more details...
Essentially, the sigma will get much smaller with large degree values, normally.
'''
```

```
mua = 1/20
mub = 1/20

def VarDeg(d):
return ( (-pa*pb/d)**2 * ((1/(mua)**4) * sigmaA**2 / Na) + (pa*pb/d) * (1/(mub)**4 * sigmaB**2 / Nb)
def VarNoDeg(d):
return ( (-pa*pb/pab)**2 * (paa*pbb / (kaa + kab)) + (pa*pb/pab)**2 * (pab*pbb / (kab + kbb)) )

def VarDegInfo(d):
return ( (-pa*pb/d)**2 * ((1/(mua)**4) * sigmaA**2 / Na) + (pa*pb/d) * (1/(mub)**4 * sigmaB**2 / Nb)
def VarNoDegInfo(d):
return ( (-pa*pb/pab)**2 * (paa*pbb / (kaa + kab)*(d-1)) + (pa*pb/pab)**2 * (pab*pbb / (kab + kbb)*(

di = np.linspace(2, 52, num = 100)
plt.figure('ExampleSimRDS')
plt.plot(di, VarDegInfo(di)/VarDeg(di), 'r', label = 'Variable degree')
plt.plot(di, VarNoDegInfo(di)/VarNoDeg(di), 'b', label = 'Set degree')
plt.xlabel('$Degree$', fontsize = 11)
plt.ylabel('Variance reduction', fontsize = 11)
plt.title('A prevalence variance comparison between including contact info and excluding contact inf
plt.legend(loc = 'upper right')
```

5.4 Mathematica Calculations of Imperfect Information

We first initialise and solve the parameters with expressions for the data for rough estimations:

```
Remove["@"]
Solve[{caa == paa*sa*na, cab == (1 - paa)*(1 - sb)*na,
cbb == (1 - paa)*sb*na} /. {na -> caa + cab + cba + cbb}, {sa, sb,
paa}]
```

Now we define the loglikelihood function:

```
l[paa_, sa_, sb_] :=
caa*(Log[paa] + Log[sa]) + cab*(Log[(1 - paa)] + Log[1 - sb]) +
cba*(Log[paa] + Log[1 - sa]) + cbb*(Log[1 - paa] + Log[sb]) +
ra*Log[paa*sa + (1 - paa)*(1 - sb)] +
rb*Log[(1 - paa)*sb + paa*(1 - sa)]
```

Taking the partial derivative to $\tilde{p}_{aa}$, after which we insert approximate values of the parameters, can be done as follows:

```
D[l[paa, sa, sb], paa] /. {sa -> caa/(caa + cba) + \[Epsilon]a,
sb -> cbb/(cab + cbb) + \[Epsilon]b,
paa -> (caa + cba)/(caa + cba + cab + cbb) + \[Epsilon]p}

Output:
cab/(1 - (caa + cba)/(caa + cab + cba + cbb) - \[Epsilon]p) - cbb/(
1 - (caa + cba)/(caa + cab + cba + cbb) - \[Epsilon]p) + caa/((
caa + cba)/(caa + cab + cba + cbb) + \[Epsilon]p) + cba/((
caa + cba)/(caa + cab + cba + cbb) + \[Epsilon]p) + (
rb (1 - caa/(caa + cba) - cbb/(
cab + cbb) - \[Epsilon]a - \[Epsilon]b))/((cbb/(
cab + cbb) + \[Epsilon]b) (1 - (caa + cba)/(
caa + cab + cba + cbb) - \[Epsilon]p) + (1 - caa/(
caa + cba) - \[Epsilon]a) ((caa + cba)/(
```

```
caa + cab + cba + cbb) + \[Epsilon]p)) + (
ra (-1 + caa/(caa + cba) + cbb/(
cab + cbb) + \[Epsilon]a + \[Epsilon]b))/((1 - cbb/(
cab + cbb) - \[Epsilon]b) (1 - (caa + cba)/(
caa + cab + cba + cbb) - \[Epsilon]p) + (caa/(
caa + cba) + \[Epsilon]a) ((caa + cba)/(
caa + cab + cba + cbb) + \[Epsilon]p))
```

We now calculate and simplify the series expansion with linearised parameters:

```
Series[\%, {\[Epsilon]a, 0, 1}, {\[Epsilon]b, 0, 1}, {\[Epsilon]p, 0,
1}]
FullSimplify[%]
```

$$\left(\left(-\frac{(caa+cab+cba+cbb)\,(-cab\,cba+caa\,cbb)\,(-(cba+cbb)\,ra+(caa+cab)\,rb)}{(caa+cab)\,(caa+cba)\,(cab+cbb)\,(cba+cbb)}+\right.\right.$$
$$(caa+cab+cba+cbb)^2\left(-\frac{caa}{(caa+cba)^2}-\frac{cba}{(caa+cba)^2}-\frac{cab}{(cab+cbb)^2}-\frac{cbb}{(cab+cbb)^2}-\right.$$
$$\left.\frac{(cab\,cba-caa\,cbb)^2\,ra}{(caa+cab)^2\,(caa+cba)^2\,(cab+cbb)^2}-\frac{(cab\,cba-caa\,cbb)^2\,rb}{(caa+cba)^2\,(cab+cbb)^2\,(cba+cbb)^2}\right)\epsilon p+O[\epsilon p]^2+$$
$$\left(\frac{(caa+cab+cba+cbb)^2\left(\frac{caa\,ra}{(caa+cab)^2}-\frac{cba\,rb}{(cba+cbb)^2}\right)}{caa+cba}+\right.$$
$$\left.\left.\frac{2\,(caa+cab+cba+cbb)^3\,(cab\,cba-caa\,cbb)\,(caa\,(cba+cbb)^3\,ra+(caa+cab)^3\,cba\,rb)\,\epsilon p}{(caa+cab)^3\,(caa+cba)^2\,(cab+cbb)\,(cba+cbb)^3}+O[\epsilon p]^2\right)\right)$$
$$\epsilon b+O[\epsilon b]^2\Bigg)+\left(\left(\left(\frac{(caa+cab+cba+cbb)^2\left(\frac{cab\,ra}{(caa+cab)^2}-\frac{cbb\,rb}{(cba+cbb)^2}\right)}{cab+cbb}+\right.\right.\right.$$
$$\left.\left.\frac{2\,(caa+cab+cba+cbb)^3\,(cab\,cba-caa\,cbb)\,(cab\,(cba+cbb)^3\,ra+(caa+cab)^3\,cbb\,rb)\,\epsilon p}{(caa+cab)^3\,(caa+cba)\,(cab+cbb)^2\,(cba+cbb)^3}+\right.$$
$$O[\epsilon p]^2\Bigg)+\left((caa+cab+cba+cbb)^2\left(\frac{(-caa+cab)\,ra}{(caa+cab)^3}+\frac{(-cba+cbb)\,rb}{(cba+cbb)^3}\right)+\frac{1}{(caa+cba)\,(cab+cbb)}\right.$$
$$2\,(caa+cab+cba+cbb)^3\left(-\frac{(-cab^2\,cba+caa^2\,(cab-cbb)+caa\,cab\,(cab+2\,(cba+cbb)))\,ra}{(caa+cab)^4}+\right.$$
$$\left.\left.\left.\frac{(cab\,cba\,(cba-2\,cbb)+caa\,cbb\,(-2\,cba+cbb)-cba\,cbb\,(cba+cbb))\,rb}{(cba+cbb)^4}\right)\right)$$
$$\epsilon p+O[\epsilon p]^2\Bigg)\,\epsilon b+O[\epsilon b]^2\Bigg)\,\epsilon a+O[\epsilon a]^2$$

Figure 4: The output for the Series command.

5.4.1 Alternative approaches

As an aside, attempting to solve the following, which involves setting all three derivatives of the loglikelihood function to the parameters to 0, results in a failure:

```
Remove["@"]
l[paa_, sa_, sb_] :=
caa*(Log[paa] + Log[sa]) + cab*(Log[(1 - paa)] + Log[1 - sb]) +
cba*(Log[paa] + Log[1 - sa]) + cbb*(Log[1 - paa] + Log[sb]) +
ra*Log[paa*sa + (1 - paa)*(1 - sb)] +
rb*Log[(1 - paa)*sb + paa*(1 - sa)]

FullSimplify[D[l[paa, sa, sb], paa]]
FullSimplify[D[l[paa, sa, sb], sa]]
FullSimplify[D[l[paa, sa, sb], sb]]
```

```
Solve[{D[l[paa, sa, sb], paa] == 0, D[l[paa, sa, sb], sa] == 0,
D[l[paa, sa, sb], sb] == 0}, {paa, sa, sb}]
```

Instead, we use the following loglikelihood function again, with estimated values for the sensitivity parameters, simplify this, and take the derivate to $\tilde{p}_{aa}$.

```
caa*(Log[paa] + Log[sa]) + cab*(Log[(1 - paa)] + Log[1 - sb]) +
cba*(Log[paa] + Log[1 - sa]) + cbb*(Log[1 - paa] + Log[sb]) +
ra*Log[paa*sa + (1 - paa)*(1 - sb)] +
rb*Log[(1 - paa)*sb + paa*(1 - sa)] /. {sa -> caa/(caa + cba),
sb -> cbb/(cbb + cab)}

Simplify[%]

D[%, paa]
```

As output, we retrieve:

```
cab (Log[cab/(cab + cbb)] + Log[1 - paa]) +
cbb (Log[cbb/(cab + cbb)] + Log[1 - paa]) +
caa (Log[caa/(caa + cba)] + Log[paa]) +
cba (Log[cba/(caa + cba)] + Log[paa]) +
ra Log[(caa paa)/(caa + cba) + (cab - cab paa)/(cab + cbb)] +
rb Log[(cba paa)/(caa + cba) + (cbb - cbb paa)/(cab + cbb)]
```

Now, as for solving the derivate to $\tilde{p}_{aa}$ by setting it equal to 0, we retrieve an extremely lengthy expression.

```
Solve[-(cab/(1 - paa)) - cbb/(1 - paa) + caa/paa + cba/
paa + ((-1 + caa/(caa + cba) + cbb/(
cab + cbb)) ra)/((1 - cbb/(cab + cbb)) (1 - paa) + (caa paa)/(
caa + cba)) + ((1 - caa/(caa + cba) - cbb/(cab + cbb)) rb)/((
cbb (1 - paa))/(cab + cbb) + (1 - caa/(caa + cba)) paa) == 0, paa]
```

We will not supply the answer, as the length of the expression amounts to about 100 pages. It is extremely lengthy, however it can theoretically be used to obtain an approximation of $\tilde{p}_{aa}$.

References

[1] Bonnie H Erickson. Some problems of inference from chain data. *Sociological methodology*, 10:276–302, 1979.

[2] Douglas D Heckathorn. Respondent-driven sampling ii: deriving valid population estimates from chain-referral samples of hidden populations. *Social problems*, 49(1):11–34, 2002.

[3] Lisa G Johnston and Keith Sabin. Sampling hard-to-reach populations with respondent driven sampling. *Methodological innovations online*, 5(2):38–48, 2010.

[4] Jon Kleinberg and David Easley. Networks, crowds, and markets: Reasoning about a highly connected world. *Cambridge University Press. C*, 1(2):3, 2010.

[5] Abdolreza Shaghaghi, Raj S Bhopal, and Aziz Sheikh. Approaches to recruiting 'hard-to-reach' populations into research: a review of the literature. *Health promotion perspectives*, 1(2):86, 2011.

[6] Margriet Spoorenberg. Estimators for respondent driven sampling, 1 2019.