Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling

Matthew J. Salganik

ABSTRACT *Hidden populations, such as injection drug users and sex workers, are central to a number of public health problems. However, because of the nature of these groups, it is difficult to collect accurate information about them, and this difficulty complicates disease prevention efforts. A recently developed statistical approach called respondent-driven sampling improves our ability to study hidden populations by allowing researchers to make unbiased estimates of the prevalence of certain traits in these populations. Yet, not enough is known about the sample-to-sample variability of these prevalence estimates. In this paper, we present a bootstrap method for constructing confidence intervals around respondent-driven sampling estimates and demonstrate in simulations that it outperforms the naive method currently in use. We also use simulations and real data to estimate the design effects for respondent-driven sampling in a number of situations. We conclude with practical advice about the power calculations that are needed to determine the appropriate sample size for a study using respondent-driven sampling. In general, we recommend a sample size twice as large as would be needed under simple random sampling.*

KEYWORDS *Design effects, Hidden populations, Power analysis, Respondent-driven sampling, Sample size, Snowball sampling, Variance estimation.*

INTRODUCTION

To understand and control the spread of HIV, it is important to have accurate information about hidden populations such as injection drug users and sex workers.[1] However, these populations are difficult to study with standard sampling methods because sampling frames do not exist. The need to gather information about such hidden populations is not limited to public health. Social scientists and policy-makers are interested in many other hidden populations such as undocumented immigrants, artists, and members of some social movements.

In response to the problem of studying hidden populations, a new statistical approach called respondent-driven sampling has been developed.[2–4] Respondent-driven sampling data are collected via a link-tracing (snowball) design, where current sample members recruit future sample members. For many years, researchers thought it was impossible to make unbiased estimates from this type of sample. However, it was recently shown that if certain conditions are met and if the appropriate procedures are used, then the prevalence estimates from respondent-driven sampling are asymptotically unbiased.[4] For example, respondent-driven sampling can be used to estimate the prevalence of HIV among drug injectors in New York City.

Salganik is with the Department of Sociology, 1180 Amsterdam Avenue, New York, NY 10027, USA.

Correspondence: Matthew J. Salganik, Department of Sociology, 1180 Amsterdam Avenue, New York, NY 10027, USA. (E-mail: mjs2105@ columbia.edu)

Despite the progress that has been made in making prevalence estimates, less is known about the sample-to-sample variability of these estimates. This gap in knowledge can lead researchers to construct inaccurate confidence intervals around estimates and to undertake studies with sample sizes that are too small to meet study goals. Filling this important gap in the respondent-driven sampling literature, this paper explores issues related to the sample-to-sample variability of estimates. The paper consists of four main parts. First, we briefly review the existing respondent-driven sampling methodology. Next, we develop and evaluate a bootstrap procedure for constructing confidence intervals around respondent-driven sampling estimates. Then, we estimate the design effect of the prevalence estimates in a number of simulated and real populations. The paper concludes with advice about the sample sizes needed for studies using respondent-driven sampling. In general, we recommend a sample size twice as large as would be needed under simple random sampling.

REVIEW OF RESPONDENT-DRIVEN SAMPLING

A respondent-driven sample is collected with a link-tracing design, similar to a snowball sample.[5–7] The sampling process begins with the selection of a set people in the target population who serve as seeds. After participating in the study, these seeds are each provided with a fixed number of unique recruitment coupons, which they use to recruit other people they know in the target population. After participating in the study, these new sample members are also provided with recruitment coupons, which they then use to recruit others. The sampling continues in this way, with subjects recruiting more subjects, until the desired sample size is reached.[2–4] Experience has shown that this sample selection method is practical and it has already been used to study a number of different hidden populations, including jazz musicians,[8] drug injectors,[2] Latino gay men,[9] and MDMA/Ecstasy users.[10]

In addition to collecting information that addresses the substantive focus of the study, a researcher using respondent-driven sampling must also collect two other pieces of information: the recruiter of each sample member and the number of relationships that each sample member has in the population of interest (sometime called degree of each person). This additional information is necessary because the estimation procedure is a variation of the traditional sampling and estimation process. Ordinarily, researchers take a sample and then use that sample to make inferences about a population (Figure 1a). However, when the data are collected with respondent-driven sampling, it is very difficult (and maybe even impossible) to make inferences from the sample directly to the population. Fortunately, such data can be used to make inferences about the social network connecting the hidden population, and this information about the social network can be used to estimate the prevalence a specific trait (Figure 1b).[3] Within this new framework, unbiased estimation becomes possible.[a] A more detail description of the estimation procedure and the conditions under which it is unbiased is available in the literature.[4]

While the ability to make unbiased prevalence estimates represented a step forward for the study of hidden populations, it was an incomplete one. In order for respondent-driven sampling to be practical as a methodology, a procedure is needed in order to put confidence intervals around these prevalence estimates.

---

[a]This new framework (Figure 1b) makes great sense when one considers the history of snowball sampling. These methods were introduced[5] and elaborated[6] to study characteristics of social networks, not the characteristics of the people in the network.[11]
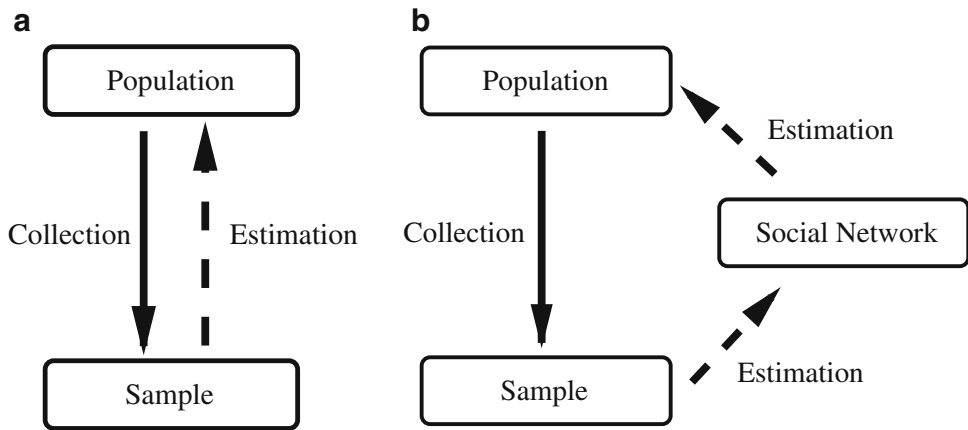
FIGURE 1. Schematic of both traditional sampling and estimation and respondent-driven sampling. By not attempting to estimate directly from the sample to the population, respondent-driven sampling avoids many of the well-known problems with estimation from snowball samples.

CONFIDENCE INTERVALS

Before introducing the confidence interval procedure, we first need to introduce some language with which to describe the hidden population. In this paper we will consider the situation of a hidden population that is made up of two mutually exclusive and collectively exhaustive groups that, for the sake of generality, we will call group A and group B. The groups could be, for example, people with and without HIV. The proportion of the population in group A will be called P_A. A point estimate of this prevalence is useful, but it is difficult to interpret without some measure of the precision of the estimate. One common way of describing this precision is with a confidence interval that provides a range within which the researcher expects to find the true population value with some level of certainty. Procedures to generate confidence intervals are well developed in the case of simple random sampling,[12,13] but researchers using a complex sample design, where not all units have the same probability of selection, are often left without guidance. Despite numerous warnings,[7,14] researchers often ignore the fact that their data were collected with a complex sample design and construct confidence intervals as if they had a random sample. This approach of ignoring the sampling design, which we will call the *naive method*, will generally cause researchers using respondent-driven sampling to produce confidence intervals that are too small. These incorrect confidence intervals are not just a technical concern; incorrect confidence intervals can lead to incorrect substantive conclusions.

In order to produce better confidence intervals, we will develop and evaluate a bootstrap method specifically designed for respondent-driven sampling.[b] Although an analytic approach would be preferable,[c] bootstrap methods are commonly used

---

[b]Some preliminary work on bootstrap procedures for respondent-driven sampling has been reported in the literature.[3] Here we build on those first steps by offering an improved procedure and a more developed analysis.

[c]We tried and failed to produce analytic results. However, some progress has been made on analytic variance estimation when an alternative estimation procedure is used.[15]

for variance estimation from complex sample designs because analytic solutions are often not possible.[16,17] In the next sections, we will describe our proposed bootstrap procedure and then evaluate its performance using computer simulations.

PROPOSED BOOTSTRAP PROCEDURE

The general idea of the bootstrap procedure is to use the observed sample to generate a set of replicate samples. Then this set of replicate samples is used to produce a set of replicate estimates. By examining the variation in these replicate estimates, one can construct a confidence interval around the original point estimate. This three-step process can be seen in Figure 2.[17]

The first step in our procedure is the resampling step. In traditional bootstrapping, this resampling is done by randomly sampling with replacement from the original sample until the replicate sample is the same size as the original sample. This resampling procedure is well grounded theoretically for the case where the original sample is collected via simple random sampling.[17] However, as described previously, in respondent-driven sampling there are dependencies in the sample selection process, and so we must use a modified resampling procedure which mimics these features. The modification of the resampling step is the main way that this approach deviates from traditional bootstrapping techniques.

Under our proposed procedure we divide the sample members into two sets based on how they were recruited: people recruited by someone in group A (which we will call A_{rec}) and people recruited by someone in group B (which we will call B_{rec}). For example, A_{rec} could be the set of all sample members who were recruited by someone with HIV. Note that this set could include both people with and without HIV. In order to mimic the actual sampling process, the resampling begins when a seed is chosen with uniform probability from the entire sample. Then, based on the group membership of the seed, we draw with replacement from either A_{rec} or B_{rec}. For example, if the seed chosen for the replicate sample was a sample member with HIV, we draw from the set of sample members who were recruited by someone with HIV. Next, we examine the group membership of this newly chosen person and then draw again with replacement from either A_{rec} or B_{rec}.[d] This process continues until the bootstrap sample is the same size as the original sample. Overall, this resampling scheme preserves some, but not all, of the dependencies that exist in the respondent-driven sampling data collection.[e]

Once the bootstrap samples are selected, we move to step 2 in Figure 2: the estimation step. Here we use the normal respondent-driven sampling estimation procedure on each of the R replicate samples to produce a set of R replicate estimates. Finally, in step 3 of the bootstrap procedure, the R replicate estimates are

---

[d]In some extremely rare cases, usually where one of the groups is very small, either A_{rec} or B_{rec} are empty. When this occurs we draw randomly from the entire sample.

[e]Simulation results indicate that this proposed procedure works better than the simpler procedure of choosing a sample member and then, based on the estimated cross-group connection probabilities, choosing a sample member from the appropriate group. The method presented here preserves those probabilities, but in addition allows for the possibility that those recruited by people in group A might be different than those recruited by people in group B.
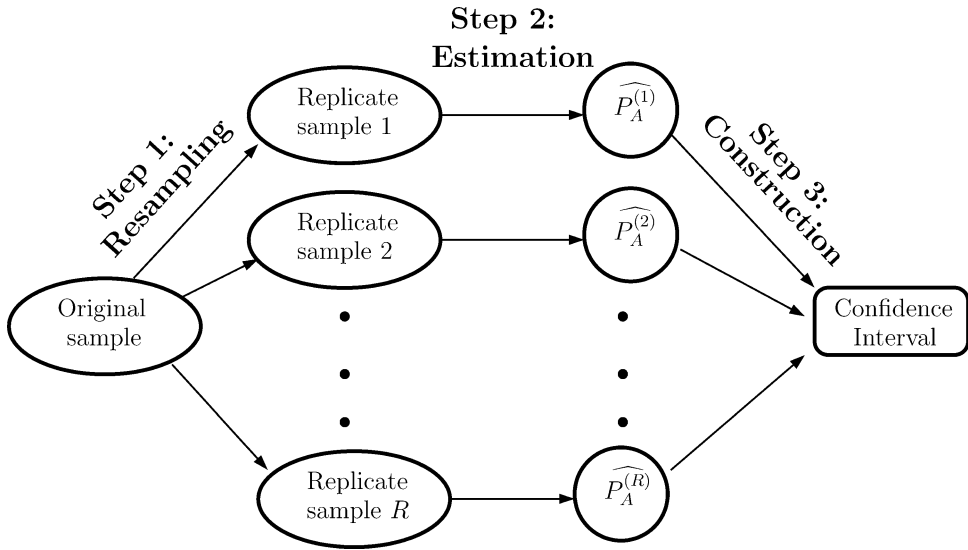
FIGURE 2. Schematic of the procedure for producing confidence intervals.

converted into a confidence interval. One way to do this would be to construct a 90% confidence interval based on the normal approximation,

$$\left\lfloor \hat{P}_A - 1.65 \cdot \hat{se}\left(\hat{P}_A\right), \hat{P}_A + 1.65 \cdot \hat{se}\left(\hat{P}_A\right)\right\rfloor \tag{1}$$

where the estimated standard error, $\hat{se}(\hat{P}_A)$, is the standard deviation of the replicate estimates. While this approach is reasonable, it has two main disadvantages. First, it forces the confidence intervals to be symmetric, which can reduce accuracy, and second, it can produce intervals with endpoints outside of the range [0, 1].

Fortunately, there are several improvements over this standard error method, and in this paper we will use the percentile method.[f] When using the percentile method, we define the endpoints of the 90% confidence interval to be the two replicate estimates, such that 5% of the replicate estimates fall below the interval, and 5% of the replicate estimates fall above the interval. For example, if a researcher generated 2,000 bootstrap replicates, a 90% confidence interval would be defined by the 100 and 1,900 ordered replicate estimates. As we shall see in the next section, the proposed resampling scheme combined with the percentile method produces confidence intervals that are generally good in an absolute sense and better than the naive method.[g]

COMPARING THE NAIVE AND BOOTSTRAP METHODS

The quality of a confidence interval procedure can be measured by calculating ϕ, the percentage of proposed intervals that contain the true population value. For

---

[f]We also attempted to use the BC_a method which, in some cases, has better asymptotic properties than the percentile method. However, in our simulations, the BC_a method performed worse. We suspect that the poor performance of the BC_a method was because of difficulties estimating the acceleration term $(\hat{a})$ when the data were collected via respondent-driven sampling.

[g]Simulations reveal that, in general, the standard error method produces intervals only slightly worse than the percentile method and so, in practice, either method can be used.

example, if we took 1,000 samples from the population and produced a 90% confidence interval from each of these samples, then 900 out of 1,000 of these confidence intervals should include the true population prevalence.[h] Unfortunately, due to resource constraints, we cannot repeatedly sample from real hidden populations. However, using computer simulations, we can construct hypothetical hidden populations and then repeatedly sample from them to evaluate the coverage properties of the different confidence interval procedures. Further, in these computer simulations we can systematically vary the characteristics of the hidden population in order to understand the effects of population and network characteristics on the quality of the proposed confidence intervals.

For example, to explore how network structure affects the quality of the confidence intervals, we constructed a series of hypothetical populations that were identical except for the amount of interconnectedness between the two groups. More specifically, we varied the ratio of the actual number of cross-group relationships to the number of possible cross-group relationship, and thus, our measure of interconnectedness, I, can vary from 0 (no connections between the groups) to 1 (maximal interconnection). All populations were constructed with 10,000 people, 30% of which were assigned a specific trait, for example HIV. Next, we began to construct the social network in the population by giving each person a number of relationships with other people in the population. The number of relationships that an individual has is called her degree. When assigning an individual's degree we wanted to roughly match data collected in studies of drug injectors in Connecticut,[2] so each person with HIV was assigned a degree drawn randomly from an exponential distribution with mean 20, and those without HIV were assigned a degree drawn from an exponential distribution with mean 10; later in this paper we will explore other degree distributions. Once the degrees were assigned, we insured that the population had the appropriate amount of interconnection between the groups.[i]

After each population was constructed, we took 1,000 samples of size 500, and for each of these 1,000 samples we constructed a confidence interval using both the naive method (i.e., ignoring the complex sample design and pretending to have a simple random sample) and the proposed bootstrap method. By seeing if each of these confidence intervals included the true population prevalence, we calculated ϕ_{naive} and ϕ_{boot}. The results of these simulations are presented in Figure 3 and reveal two important features. First, the figure shows that, for the populations used in these simulations, the proposed bootstrap procedure outperforms the naive procedure. Second, it shows that the bootstrap procedure also performs well in an absolute sense, meaning $\phi_{\text{boot}} \approx 0.9$.

To test the robustness of these findings, we explored the coverage properties in a larger portion of the possible parameter space by varying the sample size, the proportion of the population in the groups, and the average degree of the groups

---

[h]Strictly speaking, since we are sampling from a finite population we could enumerate all possible samples and then run the confidence interval procedure on every possible sample giving us the exact coverage properties of our procedure. However, the number of possible samples is astronomical, and so, following common practice, we take a sample from the set of all possible samples and use the coverage rate from these samples to estimate the true coverage rate. Thus, our presented coverage rates are only estimates of the true coverage rate with standard error, $se(\hat{\phi}) \approx \sqrt{\frac{\phi(1-\phi)}{r}} \approx \sqrt{\frac{0.9 \cdot 0.1}{1000}} \approx 0.01$. In this paper we will ignore this complication and use $\hat{\phi}$ and ϕ interchangeably.

[i]Further details about computer simulations and default parameter values can be found elsewhere.[4] Unless otherwise stated, the default parameter values were always used.
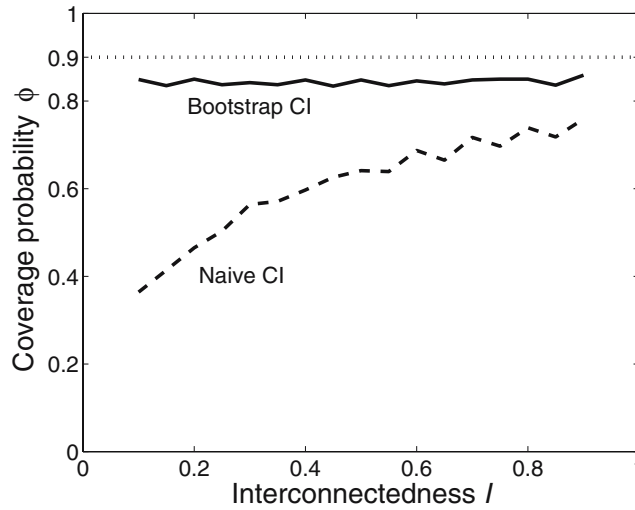
FIGURE 3. Coverage probabilities of the naive and bootstrap procedure. Results indicate that the proposed bootstrap procedure outperforms the naive procedure and performs well in an absolute sense.

(results not shown). To summarize these findings, in a few unusual portions of the parameter space, the proposed bootstrap procedure did not perform well in an absolute sense, but in most portions of the parameter space, the proposed procedure performed well.[j] Additionally, in all cases the proposed bootstrap procedure outperformed the naive procedure. To conclude, in the situations that we have examined, the proposed bootstrap procedure works well in an absolute sense and better than the naive procedure. Further, these results seem robust. Therefore, until some superior procedure is developed, we recommend this bootstrap procedure for future researchers who wish to construct confidence intervals around prevalence estimates from respondent-driven sampling.

DESIGN EFFECTS

Even though respondent-driven sampling produces unbiased prevalence estimates and allows researchers to produce approximate confidence intervals, respondent-driven sampling may not be appropriate for all situations. The unbiasedness of the estimates does not ensure that any particular estimate will be equal to the true population value. Rather, just like all other unbiased estimators, sometimes the estimate will be too low or too high; only on average will the estimate equal the true population value. If the respondent-driven sampling estimates are too variable, then, even if they are unbiased, they might not be useful in practice. For example, we constructed a population with 30% prevalence of a specific trait, for example

---

[j]The proposed bootstrap procedure performed poorly ($\phi_{\text{boot}} \approx 0.6$) when the two groups had very different total degrees ($P_A D_A \gg P_B D_B$) and I was very small ($I \approx 0.1$). As we will see in the next section, in these types of networks the design effects are large (>10), and so respondent-driven sampling probably should not be used. However, even in this extreme part of the space of all networks, the proposed bootstrap method still outperformed the naive method.

HIV, and repeatedly sampled from that population using respondent-driven sampling. To provide a benchmark for interpreting the sample-to-sample variability, we also repeatedly sampled from the population using simple random sampling (even though this would not be possible in a real hidden population). Figure 4 shows the distribution of prevalence estimates from these two methods. Both distributions center around 0.30, the true value, so both sampling methods produced unbiased estimates. But, the distribution of respondent-driven sampling estimates was more variable than the distribution of estimates under simple random sampling. A common way to quantify this difference is the design effect,[12] which measures the increased variation of the estimates under respondent-driven sampling.[k] That is,

$$deff\left(\widehat{P}_A\right) = \frac{V\left(\text{RDS},\widehat{P}_A\right)}{V\left(\text{SRS},\widehat{P}_A\right)} \qquad (2)$$

where $V\left(\text{RDS},\widehat{P}_A\right)$ is the variance of under respondent-driven sampling and $V\left(\text{SRS},\widehat{P}_A\right)$ is the variance of under simple random sampling. Because respondent-driven sampling generally provides less information, we expect that the design effect will generally, but not always, be greater than 1. For example, the results in Figure 4 show a design effect of 2.8. In the next section we will explore the design effects for a range of
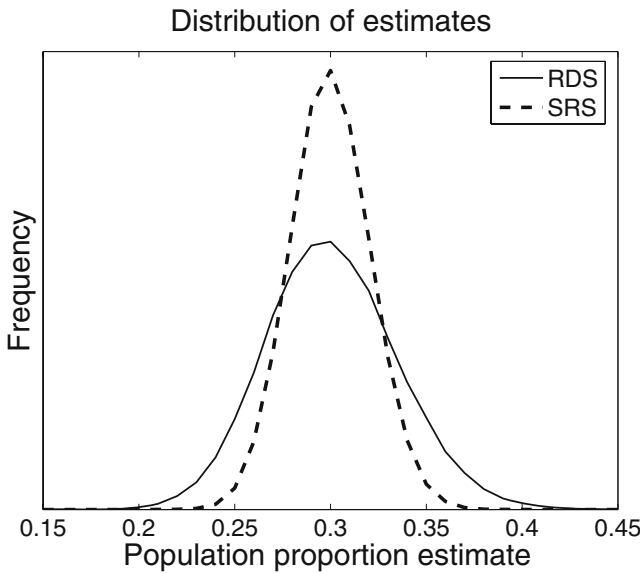

Distribution of estimates

FIGURE 4. Distribution of estimates from respondent-driven sampling and simple random sampling. The respondent-driven sampling estimates center around the true population value, 0.30, but they are more variable than the estimates from simple random sampling. This extra variation corresponds to a design effect of 2.8.

---

[k]Unfortunately, the term "design effect" has taken on two meanings in the sampling literature.[12,18] The first meaning is the ratio of the variance of the estimate under a specified sampling plan to the variance under simple random sampling (deff). An alternative definition is based on the ratio of the standard errors (deft). Since $\sqrt{deff} = deft$, readers who prefer deft can make the appropriate conversion.

different network structures to better understand the situations in which respondent-driven sampling is likely to be precise enough to be useful.

SIMULATION RESULTS ON DESIGN EFFECTS

Estimating the design effect requires comparing the variance of the prevalence estimate under different sampling methods. While the variability of the prevalence estimate under simple random sampling can be derived from basic probability theory,[12] we must use simulations to discover the variability under respondent-driven sampling. Thus, as when we evaluated the confidence interval procedure, we constructed a range of populations and simulated repeated sampling from them. We observed several general patterns that seem to occur in all portions of the parameter space. First, generally, but not always, the design effects from respondent-driven sampling were greater than 1, which indicates that respondent-driven sampling estimates were less precise than estimates from simple random sampling. This finding is consistent with the literature on complex sampling designs, which generally finds that departures from simple random sampling lead to increased variability of estimates. Second, as the interconnectedness, I, increased, that is, as the two groups became more closely connected, the design effect decreased (see Figure 5). Third, the minimum design effect for a given interconnectedness occurred not when the two groups had the same average degree ($D_A = D_B$), but when the two groups had the same total degree, ($P_A D_A = P_B D_B$) (see Figure 6). Fourth, the design effects were sensitive to the degree distribution assumed in the simulations. Previously in this paper we assumed an exponential degree distribution, but for specific subpopulations, such as drug injectors, the true functional form of the degree distribution is unknown. When we assigned a Poisson degree distribution for both groups, we observed much lower design effects, including some design effects below 1 (Figure 7); the reason for this change is currently unknown.[l] Overall, these observations should be viewed with some caution because they have not been   verified analytically due to the previously mentioned inability to develop closed-form expressions for the variance of the prevalence estimate under respondent-driven sampling.

Taken together, these simulation results suggest that the design effect is a complex function of the network structure in the population.[m] The simulation results also suggest that in some cases respondent-driven sampling can be quite blunt, with design effects as large as 10, but that in other cases it can be extremely precise, sometimes even more precise than simple random sampling.

---

[l]One possible explanation for this finding is that the Poisson distribution has lower variance than the exponential distribution; an exponential distribution has mean μ and variance μ^2, but a Poisson distribution of mean μ has variance μ.[19] However, there are also many other differences between these two distributions. To assess the role of the variance in the degree distribution on the design effects, we ran simulations where we assigned both groups a normal degree distributions. In this case, direct manipulation of the variability of the degree distribution did not change the estimated design effect.

[m]The complicated relationship between network structure and design effects implies that the relationship between homophily[2,3] and design effect is many-to-many. That is, many homophily values yield the same design effect, and a given design effect is consistent with many different homophily values. Therefore, homophily is not the best way to understand design effects.
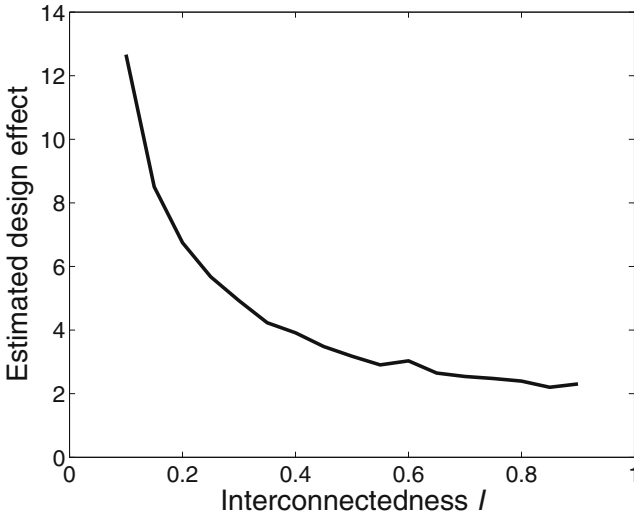
FIGURE 5. Design effect as a function of interconnectedness I. In general, as the interconnectedness increases the design effect decreases. Results are based on 10,000 replicate samples.

ESTIMATED DESIGN EFFECTS IN REAL STUDIES

The simulation results indicate that a range of design effects are possible. Therefore, an important question becomes: What are the design effects in populations that people actually study? Our best attempt to answer that question is presented in Table 1, where we report the estimated design effects from all studies that are
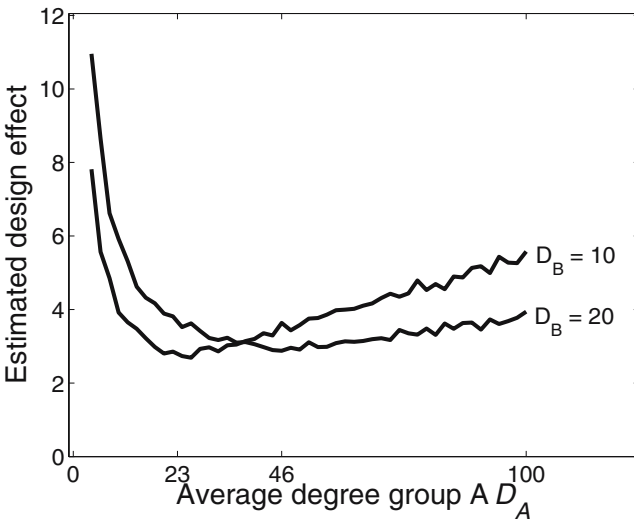

FIGURE 6. Design effect as a function of D_A for different D_B. In general, the minimum design effect, for a given interconnectedness, occurs when the two groups have the same total degree ($P_A D_A = P_B D_B$). So if $P_A = 0.3$ and $P_B = 0.7$, then when $D_B = 10$ the minimum design effect occurs when $D_A \approx 23$ and when $D_B = 20$ the minimum occurs when $D_A \approx 46$. Results are based on 10,000 replicate samples.
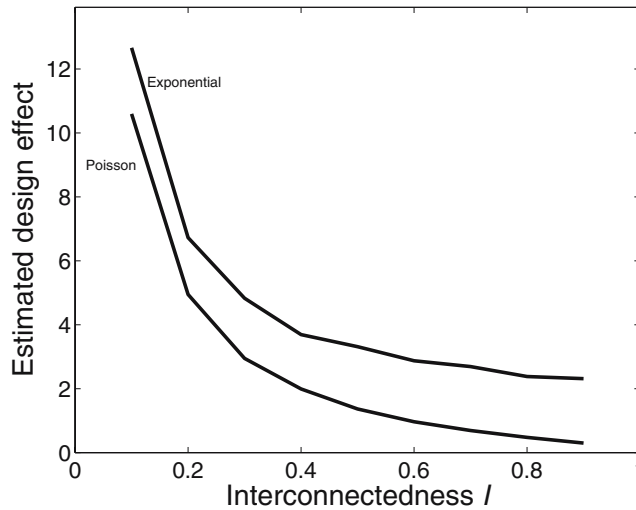
FIGURE 7. Design effect as a function of interconnectedness for different degree distributions. In general, the design effects are smaller when the degree distribution of the groups is Poisson rather than exponential. Results are based on 10,000 replicate samples.

currently available.[n] To produce the estimated design effects we took the published estimates of P_A and used them to estimate the variability of the prevalence estimates $(\widehat{V}(\mathrm{SRS}, \widehat{P}_A))$. This variability is then compared to the published estimates of the variability under respondent-driven sampling $(\widehat{V}(\mathrm{RDS}, \widehat{P}_A))$.[o] We report only one design effect because, due to the symmetry of the two-group system, $deff(\widehat{P}_A) = deff(\widehat{P}_B)$.

Overall, Table 1 shows that the prevalence estimates from existing studies had design effects around 2, suggesting that respondent-driven sampling is reasonably precise in the situations in which it has been used so far.[p] Based on this crude analysis of existing respondent-driven sampling data, we recommend that when planning a study using respondent-driven sampling researchers should assume a design effect of 2. This guideline should only be considered a preliminary rule-of-thumb and should be adjusted, if necessary, depending on pre-existing knowledge of the study population.

SAMPLE SIZE CALCULATION

Information on design effects should be used when planning the sample size of a study using respondent-driven sampling, or else the sample size will not meet the goals of the study. Fortunately, once the researcher has an estimated design effect, it is rather straightforward to adjust the required sample size; the researcher need only to multiply the sample size needed under simple random sampling by the assumed

---

[n]The results presented here for Latino gay men differ from the results originally published[9] because the standard errors published in the original paper were too large (D. Heckathorn, [ddh22@cornell.edu], email, February 5, 2006).

[o]Since these authors all used the bootstrap procedure proposed in this paper, their confidence intervals allow reasonable estimation of $\widehat{V}(\mathrm{RDS}, \widehat{P}_A)$.

[p]Table 1 also illustrates that the design effect is a property of a specific estimate and not a population; estimates from the same dataset may have different design effects.[12]

TABLE 1. Estimated design effects from real respondent-driven sampling data.[8–10] Sample sizes vary within the same study due to missing data

Study description				Study results			
Population	Location	n	Trait	$\widehat{P}_A$	$\widehat{V}\left(\text{RDS}, \widehat{P}_A\right)$	$\widehat{V}\left(\text{SRS}, \widehat{P}_A\right)$	$\widehat{\text{deff}}\left(\widehat{P}_A\right)$
Latino gay men	Chicago	69	HIV+	0.17	0.0024	0.0021	1.1
Latino gay men	San Francisco	72	HIV+	0.49	0.0041	0.0035	1.2
MDMA/ Ecstasy users	Ohio	374	Male	0.58	0.0012	0.0007	1.7
Jazz musicians	New York City	263	Male	0.76	0.0016	0.0007	2.3
Jazz musicians	New York City	261	Union member	0.25	0.0010	0.0007	1.4
Jazz musicians	New York City	253	Received airplay	0.75	0.0017	0.0007	2.4

design effect. Thus, for studies using respondent-driven sampling we recommend a sample size twice as large as would be needed under simple random sampling. However, calculating the appropriate sample size under simple random sampling is often difficult due to the overly general nature of the power analysis literature.[20,21] Therefore, we will review the sample size calculations for two specific cases of most interest to researchers using respondent-driven sampling: estimating the prevalence of a trait with a given precision and detecting a change in prevalence over time.[q]

One common goal of studies is to estimate the prevalence of a characteristic with some pre-specified precision, for example, to estimate the proportion of sex workers in New York City that are HIV-positive with a standard error of no greater than 0.03. Since it is the case that,

$$V\left(\hat{P}_A\right) = deff \cdot \frac{P_A(1 - P_A)}{n} \tag{3}$$

we can solve for the required sample size, n, in terms of the desired standard error, which yields,

$$n = deff \cdot \frac{P_A(1 - P_A)}{\left(se\left(\hat{P}_A\right)\right)^2} \tag{4}$$

Therefore, if based on pre-existing knowledge we suspect that 20% of the sex workers have HIV and that the design effect is 2, we would need a sample size of at least 356 sex workers to estimate the HIV prevalence with a standard error no greater than 0.03. Notice that this calculation depends on our initial guess of the prevalence. If researchers do not have enough information to make such a guess, they should assume a value of 0.5 which is maximally conservative.

---

[q]In addition to making prevalence estimates, some researchers are interested in using statistical techniques like multivariate regression to look for statistical patterns within the data. The feasibility of this approach is discussed elsewhere.[22]

A second problem of interest to many researchers is comparing the prevalence of some behavior at two time points. For example, a researcher might want to test whether an outreach program was successful at getting drug injectors to stop sharing injection equipment. Assume that the researcher suspects that before the intervention 40% of drug injectors share injection equipment and that the researcher would like to choose the appropriate sample size to be able to detect a drop to 25% such that there is an 80% probability that a 95% confidence interval for the estimated difference will not include 0. Further, assume that the researcher suspects that each prevalence estimate will have a design effect of 2. Based on a derivation available in the literature,[23] we can calculate that the required sample size is,

$$n = 2 \cdot \frac{\left[2.8 \cdot \sqrt{0.6 \cdot 0.4 + 0.25 \cdot 0.75}\right]^2}{(0.4 - 0.25)^2} = 298 \text{ (in each sample)} \qquad (5)$$

More generally, the required sample size for comparing prevalence in two populations is,

$$n = \text{deff} \cdot \frac{\left[\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right) \cdot \sqrt{P_{A,1}\left(1 - P_{A,1}\right) + P_{A,2}\left(1 - P_{A,2}\right)}\right]^2}{\left(P_{A,2} - P_{A,1}\right)^2} \text{ (in each sample) } (6)$$

where $Z_{1-\frac{\alpha}{2}}$ and $Z_{1-\beta}$ are the appropriate values from the standard normal distribution and *deff* is the design effect.[r]

These sample size calculations are based on assumptions about the prevalence of the characteristics and the design effect. Therefore, the sample sizes produced by Eqs. 4 and 6 should be considered approximate.

CONCLUSIONS

This paper makes two main contributions to the literature on respondent-driven sampling. First, we introduce a bootstrap confidence interval procedure that in simulations outperforms the naive method currently in practice. Therefore, we recommend this bootstrap procedure be used in future analysis of respondent-driven sampling data. The procedure requires some custom computer programming to implement, but, fortunately, it is already included in RDSAT, a software package for organizing and analyzing respondent-driven sampling data.[s]

The second major contribution of this paper is the information on design effects. The simulation results suggest that the design effects can range from as high as 10 to less than 1. These findings imply that, because of the possibility of high design effects, respondent-driven sampling is not appropriate in all cases. In some extreme network structures, the prevalence estimates could be so variable that, even though they are unbiased, they might not be very useful. Fortunately, data from existing studies suggest that, so far, respondent-driven sampling has been used in situations where it is

---

[r]This formula is an approximation of the more complicated formula derived elsewhere,[24]:
$n = \text{deff} \cdot \frac{\left[Z_{1-\frac{\alpha}{2}} \cdot \sqrt{2\overline{P}(1-\overline{P})} + z_{1-\beta} \cdot \sqrt{P_{A,1}(1-P_{A,1}) + P_{A,2}(1-P_{A,2})}\right]^2}{(P_{A,2} - P_{A,1})^2}$ where $\overline{P} = \frac{P_{A,1} + P_{A,2}}{2}$, which has appeared in the public health literature.[25]
When $P_{A,1} \approx P_{A,2}$ then $2\overline{P}(1 - \overline{P}) \approx P_{A,1}(1 - P_{A,1}) + P_{A,2}(1 - P_{A,2})$ so both formula yield similar values.

[s]The RDSAT software was written by Erik Volz and Doug Heckathorn and is currently available from http://www.respondentdrivensampling.org.

reasonably precise, yielding estimated design effects around 2 (see Table 1). Based on these data, we suggest that when using respondent-driven sampling, researchers collect a sample twice as large as would be needed under simple random sampling.

The sensitivity of the design effect to the functional form of the degree distribution further emphasizes the need for more research on methods to accurately measure the degree of each respondent. Currently, the estimated average degree depends on subjects' self-reported degree, and these reports may be inaccurate.[26,27] In almost all cases, inaccuracy in the self-reported degree will introduce bias into the prevalence estimates.[4] As far as we know, the best methods for estimating an individual's degree are scale-up method and summation method.[28] However, it is not clear that either of these approaches, which were designed for the general population, is appropriate for studying hidden populations.

Taken together, the results about the sample-to-sample variability presented in this paper add to the growing literature on respondent-driven sampling. By allowing researchers to obtain better information about key hidden populations, this research should allow public health professionals to monitor population dynamics more accurately, target resources more carefully, and intervene to slow the spread of disease more effectively.

ACKNOWLEDGEMENTS

REFERENCES

1. Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*. 2005;19(Supp 2):S67–S72.
2. Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. *Soc Probl*. 1997;44(2):174–199.
3. Heckathorn DD. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Soc Probl*. 2002;49(1):11–34.
4. Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol Method*. 2004;34:193–239.
5. Coleman JS. Relational analysis: the study of social organization with survey methods. *Human Organ*. 1958;17:28–36.
6. Goodman L. Snowball sampling. *Ann Math Stat*. 1961;32(1):148–170.
7. Thompson SK, Frank O. Model-based estimation with link-tracing sampling designs. *Sur Methodol*. 2000;26(1):87–98.
8. Heckathorn DD, Jeffri J. Finding the beat: using respondent-driven sampling to study jazz musicians. *Poetics*. 2001;28:307–329.
9. Ramirez-Valles J, Heckathorn DD, Vázquez R, Diaz RM, Campbell RT. From networks to populations: the development and application of respondent-driven sampling among IDUs and Latino Gay Men. *AIDS Behav*. 2005;9(4):387–402.
10. Wang J, Carlson RG, Falck RS, Siegal HA, Rahman A, Li L. Respondent-driven sampling to recruit MDMA users: a methodological assessment. *Drug Alcohol Depend*. 2005;78:147–157.

11. Snijders TAB. Estimation on the basis of snowball samples: how to weight? *BMS Bull Méthodol Sociol*. 1992;36:59–70.
12. Lohr SL. *Sampling: Design and Analysis*. Pacific Grove: Duxbury; 1999.
13. Thompson SK. *Sampling*. New York: Wiley; 2002.
14. Thompson SK, Collins LM. Adaptive sampling in research on risk-related behaviors. *Drug Alcohol Depend*. 2002;68:S57–S67.
15. Volz E, Heckathorn DD. Probability-based estimation theory for respondent-driven sampling. *Working Paper*. 2006.
16. Wolter KM. *Introduction to Variance Estimation*. Berlin Heidelberg New York: Springer; 1985.
17. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall; 1993.
18. Lu H, Gelman A. A method for estimating design-based sampling variances for surveys with weighting, poststratification, and raking. *J Off Stat*. 2003;19(2):133–151.
19. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Boca Raton: Chapman & Hall; 2004.
20. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1987.
21. Murphy KR, Myors B. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. Mahwah, NJ: Lawrence Erlbaum Associates; 1998.
22. Heckathorn DD. Extensions of respondent-driven sampling: dual-components sampling weights. Paper presented at: RAND Statistical Seminar Series, 2005; Santa Monica, CA.
23. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel / Hierarchial Models*. Cambridge: Cambridge University Press; 2006.
24. Fleiss JL. *Statistical Methods for Rates and Proportions*. New York: Wiley; 1973.
25. FHI. *Behavioral Surveillance Surveys: Guidelines for Repeated Behavioral Surveys in Populations at Risk of HIV*. Arlington, VA: Family Health International; 2000.
26. Brewer DD. Forgetting in the recall-based elicitation of personal and social networks. *Soc Netw*. 2000;22:29–43.
27. Bell DC, Belli-McQueen B, Haider A. Partner naming and forgetting: Recall of network members. *Working Paper*. 2006; (in press).
28. McCarty C, Killworth PD, Bernard HR, Johnsen EC, Shelley GA. Comparing two methods for estimating network size. *Human Organ*. 2001;60(1):28–39.