The Relative Statistical and Operational Plausibility of Multiple Frame Sampling for Rare Population Subgroups

William D. Kalsbeek[1], Bruce D. Spencer[2], and Carol C. House[3]

Abstract

For studies where the objective is to estimate the prevalence rate of members of a sampled population who fall in a rare subgroup, this paper examines the relative statistical precision of prevalence estimates from a multiple frame sample design as compared to a single frame household design with the same data collection budget. We initially examine relative cost-efficiency for simple un-clustered samples and then briefly consider the effect of cluster sampling. Findings are illustrated for the case where the subgroup consists of victims of rape and sexual assault (RSA) within the civilian non-institutionalized population of persons 12 years and older. Two sample designs are considered: (i) dual frame sampling from a conventional household frame plus a frame constructed from police reports of RSA, versus (ii) single frame sampling from the household frame. We conclude for this simple illustration that a dual frame design will be more cost-effective in estimating RSA prevalence, to the extent that this prevalence among police reports exceeds the prevalence in the population as a whole. However, gains in the dual frame design are diminished in direct relationship to the size of intra-class correlation when cluster sampling is considered.

Key Words: multiple frame sampling; sample design; cost-efficiency

1. Introduction

Devising practical methods for sampling rare populations remains an important topic on the frontier of survey statistics. One challenge in this area is to estimate the prevalence rate for members of a rare subgroup in a more general population. That population may be sampled using a general-purpose design. This paper specifically focuses on the relative statistical and operational plausibility of enhancing a general purpose, single frame design with a multiple frame approach in which the additional frame(s) would have a substantially higher prevalence rate for members of the targeted subgroup. The authors developed this paper because of their interest in improving the estimation of the number of rapes and sexual assaults (RSA) occurring in the United States, as measured in the National Crime Victimization Survey[4] (National Research Council, 2013). In this paper,

[1] Department of Biostatistics, University of North Carolina at Chapel Hill; bill_kalsbeek@.unc.edu

[2] Department of Statistics and Institute for Policy Research, Northwestern University; bspencer@northwestern.edu

[3] Committee on National Statistics, National Academies of Science; chouse@nas.edu

[4] The National Crime Victimization Survey (NCVS) is a national household survey, conducted by the Census Bureau for the Bureau of Justice Statistics (BJS). It provides statistics on a wide range of criminal victimizations including rape, robbery, assault, theft, household burglary and motor theft. Its target population is the civilian non-institutionalized population of the U.S. who are 12 years old and older.

the authors look at a simplified example of an alternative multiple frame approach to this problem in order to assess potential feasibility.

Definitions used in this paper:

> Subgroup: members of a population possessing some particular attribute of interest.
> Rare Subgroup: a subgroup representing a very small proportion of the population (e.g., < 1%).
> *P:* proportion of the subgroup within the larger population.

2. Approach

A standard approach to sampling a general population within the U.S. is to select a sample of household residences from an area frame ("HH frame") and subsequently to list their household members. A common strategy used to estimate characteristics of a subgroup of the population is to define the sampling strata so as to isolate higher concentrations of the subgroup in these strata, and then to disproportionately sample (oversample) those strata with relatively high subgroup concentrations. This approach has several limitations. First, it may be difficult to isolate the targeted subgroup via strata groupings. Second, such a disproportionate sample of the general population may not be efficient if the survey also needs estimates of the general population or other subgroups.

The multiple frame approach discussed in the paper would combine the general-purpose household (HH) frame but also incorporate one or more administrative lists. These administrative lists would be expected to contain a higher percentage of the rare subgroup compared to the general population. Like the high concentration strata in the HH frame, these administrative frames would be oversampled for the targeted subgroup. In fact, continuing to oversample the sampling strata in the HH frame may also be prudent. A complication in the multiple frame approach is that the HH frame and the administrative frames may overlap. This complication can be circumvented through screening. Alternatively, it could be accommodated using existing estimation theory for multiple frame sampling and estimation (Hartley, 1962; Lohr, 2011).

The most useful supplementary frames for estimating the subgroup prevalence rate would be administrative lists from organizations seeing a notably higher concentration of the subgroup. Examples of possible administrative frames that could be used to measure incidences of rape and sexual assault include (but are not limited to) the following:

- Police records for individuals filing a claim of assault,
- Emergency room records for those treated for assault-related injuries,
- Records from victim support service providers for rape and domestic violence, and
- Indian Health Service records.

In the simplified scenario provided in this paper, the authors use a single administrative list (police records) combined with the HH Frame as the multiple frame for discussion.

3. Comparison

This paper compares the sampling error efficiency of the estimate of the proportion (*P*) of victims of rape and sexual assault in the general population. We simplified this comparison in several ways. First, we ignored non-sampling errors. Second, we constructed the multiple frame with only one administrative frame, so the comparative alternatives are the single frame (HH Frame) versus a dual frame design. Third, we handled the complication of frame overlap by presuming that the members of the administrative frame could be screened out of the HH Frame, thus creating two non-overlapping strata. Finally, we assume *SRSWR* sampling (of the HH Frame, and within each strata of the dual frame) that is both unclustered and unstratified. In summary:

Dual Frame = HH Frame + Police Records

Uses a stratified *SRSWR (n_h* of *N_h)* sample with "optimum" allocation of the sample between the following two strata.

Stratum A: *SRSWR* of persons filing assault claims with the police
Stratum B: *SRSWR* of persons in the HH frame who did not file a police claim of assault (screening required to make this determination)

Single Frame = HH Frame

Uses an unstratified *SRSWR* (*n* of *N*).

3.1 Optimum Strata Allocation in the Dual Frame

It is expected that sampling, recruitment, and interviewing from the police records administrative list will be substantially more difficult, and thus more costly, than for the single frame. The question is what allocation of the sample should be made between the screened HH sample and the police records strata to make the dual frame sample most cost-efficient. To answer this question, we applied the conventional theory of optimum stratum allocation with *SRSWR* selection in each stratum, where

$W_h = N_h / N$ for the *h-th* stratum, and

$P_w = \sum_{h=1}^{H} W_h p_h$ is the combined estimator of *P*.

The variance model for this case is

$$V(P_w) = \sum_{h=1}^{H} W_h^2 P_h (1-P_h) / n_h .$$

The assumed cost model is

$$\text{Cost } = \sum_{h=1}^{H} C_h n_h ,$$

where C_h = the average per-unit cost in the *h-th* stratum.

The optimum size for samples from the police records administrative stratum and the screened HH stratum is generally given by

$$n_h^{(opt)} \propto W_h \sqrt{P_h(1-P_h)} \,/\, \sqrt{C_h}\,.$$

Thus the optimum sample size for the *H=2* strata in this comparison are defined as:

- Administrative (police records) stratum A: $n_A^{(opt)}$
- Screened (non-administrative) household stratum HH: $n_{HH}^{(opt)}$

It is important to note that the relative sizes of the optimum values for the two strata in this simplified dual frame (DF) design will depend on the parameter θ, which is defined as the ratio of the average per-unit cost for the administrative stratum to the corresponding per-unit cost for the screened HH stratum. That is,

$$\theta = C_A \,/\, C_{HH}\,.$$

Based on the optimum allocation to the two strata one can determine the standard error (*SE*) of the combined estimate of *P*, $SE(p_{w,DF}) = \sqrt{V(p_{w,DF})}.$ One can also determine the expected number of subgroup members (victims of rape or sexual assault) in the dual frame sample, $P_A n_A^{(opt)} + P_{HH} n_{HH}^{(opt)}$, where P_A and P_{HH} are the subgroup prevalence rates in the screened administrative and household strata, respectively.

The count of the expected number of subgroup members is important in this comparison because, in addition to estimating *P*, data analysts are often interested in examining this subgroup of RSA victims to learn about treatment, coping strategies, and other subsequent manifestations of victimization. The statistical quality of these supplementary analyses is at least partially determined by the respondent count in the subgroup -- the larger the better. The amount of sample disproportionality created by the optimum allocation also affects the statistical quality of estimates in these analyses of victims.

3.2 Quality and Sample Size of the Single Frame

For the single frame (SF) alternative design, the comparable measures of the precision of estimates of *P* and the expected number of subgroup members are provided. The cost-comparable sample size for the single frame is:

$$n_{SF} = Cost \,/\, C_{HH}\,.$$

The standard error of the estimate of *P* is:

$$SE(p_{SF}) = \sqrt{P(1-P)/n_{SF}}\,.$$

The comparable expected number of subgroup members in the sample will be PN_{SF}.

4. Results

In this section, we present simulated results that are dependent on the parameter θ, the ratio of the average unit cost for the administrative frame to the corresponding unit cost for the screened HH frame. To run the calculations, the authors used the following rounded values for different paramenters:

- Size of target population: N = 250 million (2007 U.S. population 12+ years old)
- Size of the police frame: N_A = 140,000 (extrapolation from UCR) [5]
- Anticipated value of rape and sexual assault victimization rate: P = 0.001 [6]
- Anticipated value of rape and sexual assault victimization that were reported to the police: P_A = 0.83 = 116,000 / 140,000 [7]
- The data collection budget for the NCVS: *Cost* = $26 million [8]
- The average cost of sampling, recruiting, and interviewing respondent on the HH frame : C_{HH} = $173 = $26 million / 150,000.

The following graphics show the percent reduction in the standard error (*SE*) of the estimate of *P* for the dual frame design compared to the single frame design. The figures also show the percent increase in the number of expected subgroup members (victims of rape and sexual assault) discovered in the dual frame sample compared to the single frame sample.

The horizontal axis in each graph is the value of P_A, the proportion of the members of the screened administrative stratum that were victims of rape and sexual assault (RSA). Note that the frame from which this stratum's sample is drawn consists of individuals who reported some type of assault to police, but not necessarily rape and sexual assault. Also note that we consider the most plausible value of P_A in each of these figures to be 0.83.

Plotted values on the vertical axis provide an indication of the robustness of the indicted reduction in SE or the increase in subgroup sample sizes.

[5] UCR is the Uniform Crime Reporting System of the FBI. http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/1997/toc97.pdf

[6] From the Bureau of Justice Statistics, 2008. http://bjs.ojp.usdoj.gov/content/pub/pdf/cv07.pdf

[7] From NCVS, 1992 – 2000. http://bjs.ojp.usdoj.gov/content/pub/pdf/rsarp00.pdf

[8] FY2009 cost of the NCVS with sample size = 150,000. www.fcsm.gov/09papers/**Rand**_X-B.doc.

4.1 Results for $\theta = 2$

The graphics in this section provide results for the assumption that the unit cost of sampling, recruitment, and interviewing members of the administrative frame is 2 times the unit cost of the screened HH frame ($173).

Figure 1 shows the percent reduction in the SE for $\theta = 2$. One sees that with a fixed cost, the most cost-efficient stratum allocation for the dual frame results in a 25.9% reduction in the SE when $P_A = 0.83$. The gain in precision goes down with the value of P_A, but the graph shows that even with P_A as low as 0.50, the dual frame realizes at least a 15% drop in SE compared to the single frame.

FIGURE 1
Percent Reduction in Standard Error of Dual Frame Estimate of *P* by Proportion of RSAs in Administrative Frame When Theta=2

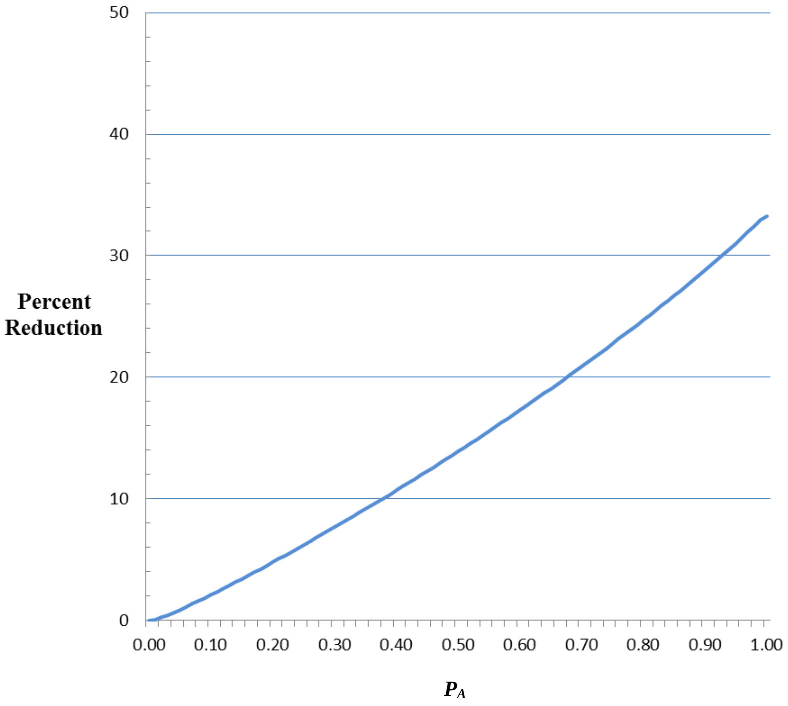
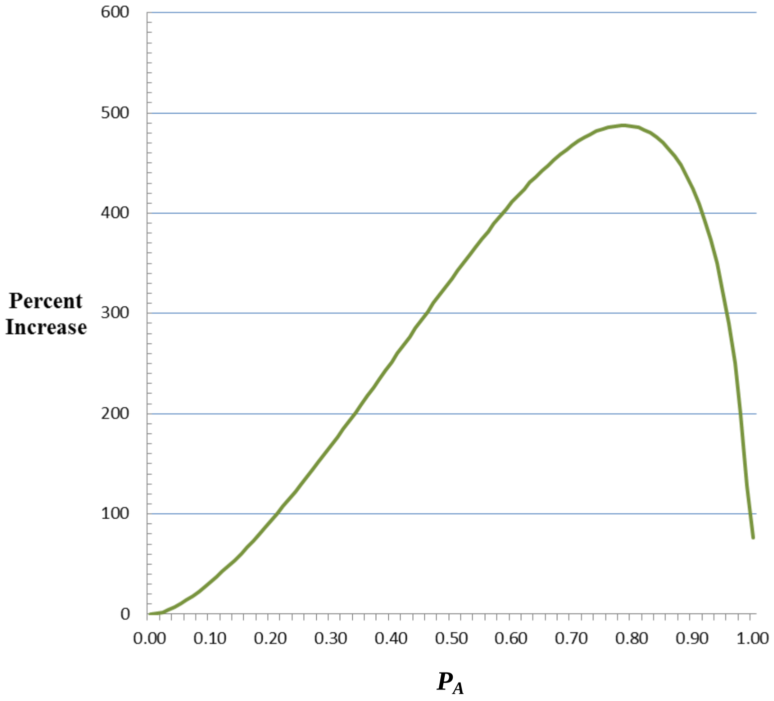


Figure 2 shows the percent increase in the expected number of discovered subgroup members (RSA victims) for $\theta = 2$. Here one can note that at the most plausible value of P_A, the dual frame found 872 subgroup members, yielding nearly six times the number of subgroup members found by the single frame (150). When P_A is as low as 0.20, the dual frame is still producing at least double the number of subgroup members as is the single frame.

FIGURE 2
Percent Increase in Number of Respondents Who Are RSA Victims by Proportion of RSAs in Administrative Frame When Theta=2



4.2 Results for $\theta = 10$

In this section the paper examines the results when there is an even greater cost differential between working the administrative frame as compared to the household frame. The graphs are similar to those in section 4.1, except that $\theta = 10$, assuming that the unit cost of sampling, recruitment, and interviewing members of the administrative frame is 10 times the unit cost of the screened HH frame ($173).

Figure 3 shows that the gains in precision of estimates of P from the optimized dual frame is still substantial. At the most plausible value of P_A the percent reduction in the SE is still around 25%. This means that the precision gains don't appear to be very sensitive to somewhat greater difficulty and cost in working with this administrative frame. This may be because the optimum sample sizes for the administrative frame (police records) are relatively small compared to the sample size from the screened HH frame for both $\theta = 2$ and $\theta = 10$.

FIGURE 3
Percent Reduction in Standard Error of Dual Frame Estimate of *P* by Proportion of RSAs in Administrative Frame When Theta=10

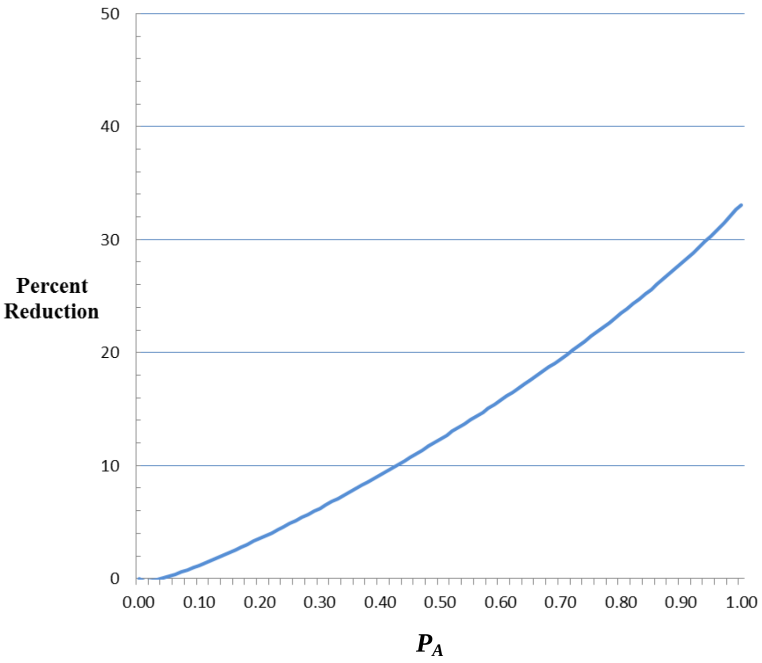
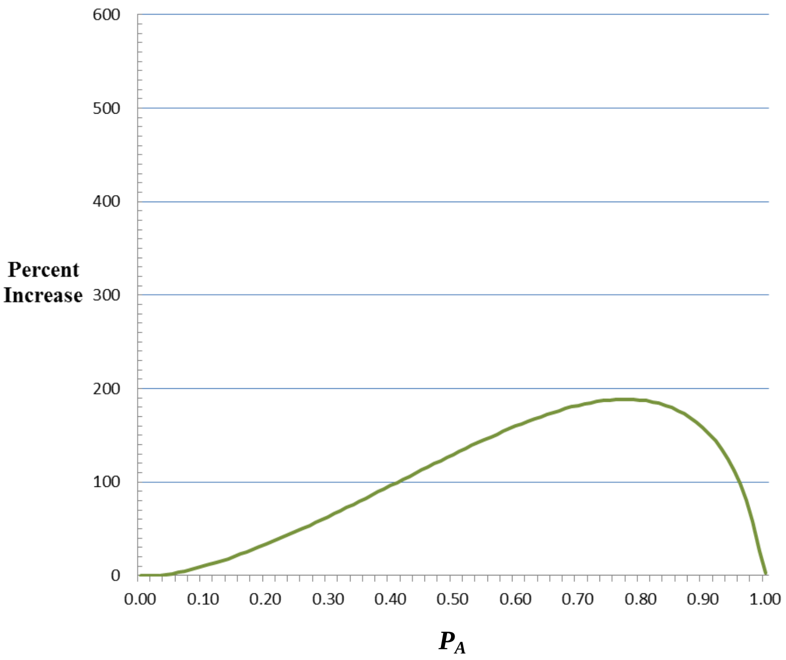


Figure 4 shows the percent increase in the expected number of discovered subgroup members (RSA victims) for θ *=10*. In contrast to the change in precision, the change in the number of discovered RSA victims in the dual frame falls off considerably when the simulation moves from θ *= 2* to θ *=10*, but there are still gains. At the most plausible value of P_A the analysis shows about three times the number of RSA victims in the dual frame compared to the single frame approach (427 versus 150).

FIGURE 4
Percent Increase in Number of Respondents Who Are RSA Victims by Proportion of RSAs in Administrative Frame When Theta=10



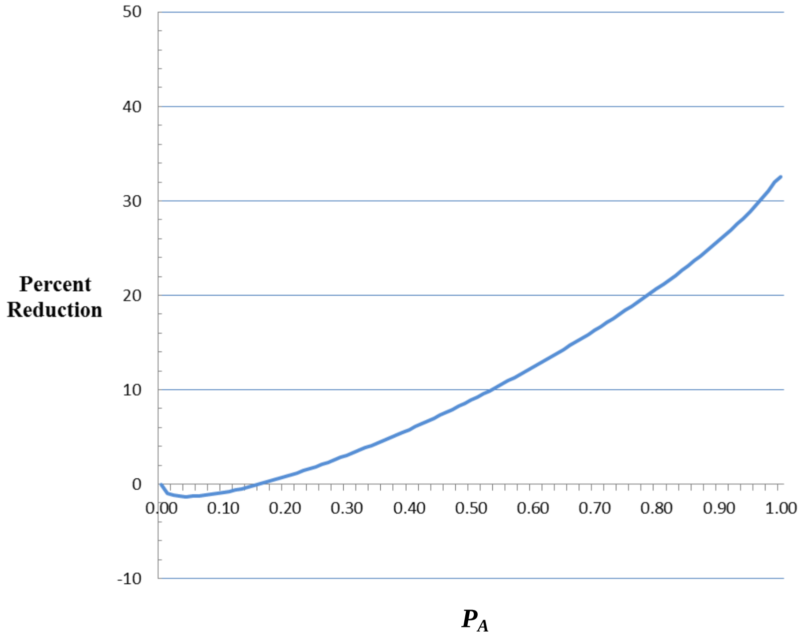
4.3 Results for θ = *50*

This simulation concludes with an examination of results when the per-unit cost differential between the administrative and household frames is 50 times higher. The graphs are similar to those in sections 4.1 and 4.2, except that θ = *50*.

Figure 5 shows the gains in precision of estimates of *P* from the optimized dual frame is still substantial, although gains are somewhat less that for smaller values of θ. The percent reduction in the SE at the most plausible value of P_A is down to 22 percent. When P_A is above 0.50, then the dual frame continues to produce a 10% or greater increase in precision. However, one can also note that if the actual value of P_A were less than about 0.15, the single frame design produces more precise estimates.

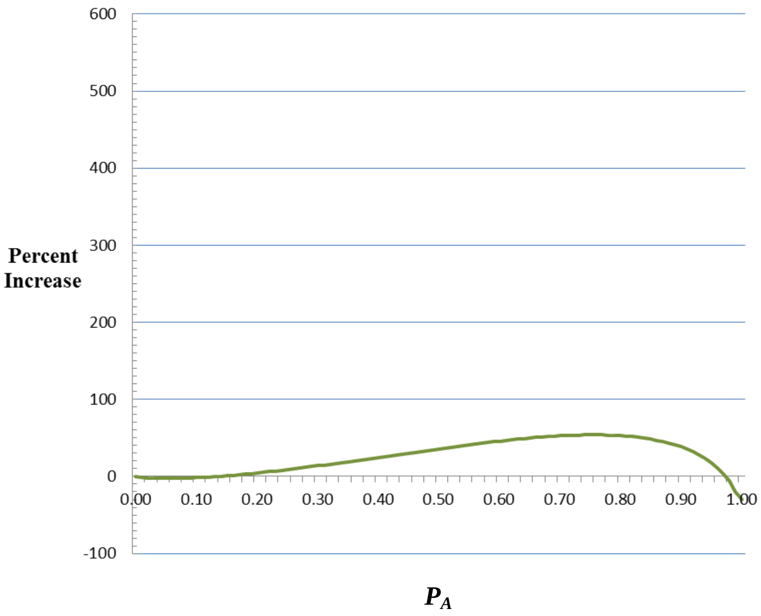
FIGURE 5

Percent Reduction in Standard Error of Dual Frame Estimate of *P* by Proportion of RSAs in Administrative Frame When Theta=50



In the final graphic, Figure 6, the results for the percent increase in the expected number of discovered subgroup members (RSA victims) for θ *= 50* are substantially less than for lower values of θ. At the most plausible value of P_A, the dual frame would find about 50 percent more RSA victims. The non-monotonic shape of the plot in this figure (as well as in previous comparable figures) is primarily due to the fixed-cost constraint we established for these comparisons and to the non-linear relationship between P_A and $P_A(1-P_A)$ across the range of P_A.

FIGURE 6
Percent Increase in Number of Respondents Who Are RSA Victims by Proportion of RSAs in Administrative Frame When Theta=50



5. Efficiency Reductions from Cluster Sampling

Let the variance of the estimate of RSA have the form (from single frame sampling) $V_{SF}(\hat{P}) = V_1 + V_2$, where $V_1 > 0$ denotes the component of variance from all stages of sampling except the final stage, and where $V_2 > 0$ denotes the variance from the final stage of sampling. The effect of using dual frame sampling is to reduce the variance to $V_{DF}(\hat{P}) = V_1 + (1-g_2)V_2$, where g_2 denotes the variance reduction from element sampling. Figure 1 suggests that $g_2 \approx .259^2 = .0671$ when the percentage of RSA victims in the administrative frame is 83%, or $P_A = 0.83$. Note that g_2 does not tell us the overall variance reduction. The overall variance reduction attributable to dual frame sampling, g, satisfies $V_{DF}(\hat{P}) = (1-g)(V_1 + V_2)$. It follows that $g = g_2 / (1 + V_1 / V_2) < g_2$. Now consider the special case of two-stage sampling with equal size clusters in the single frame sample, SRS at both stages. Let b denote the number of selections per PSU in the single frame sample. Let the intra-class correlation for the quantity of interest (RSA) be $\rho = \sigma_1^2 / (\sigma_1^2 + \sigma_2^2)$, where we ignore finite population considerations. Then the design effect, $\delta,$ is approximately equal to $1+(b-1)\rho.$ Note that $V_1/V_2 \approx b\sigma_1^2/\sigma_2^2$ $\approx b\rho/(1-\rho)$ and thus

$$\frac{g_2}{g} \approx 1+\frac{b\rho}{1-\rho} \approx \frac{\delta}{1-\rho}.$$

6. Conclusions and Future Research

This paper presented a comparison of two simplified approaches (single frame versus dual frame) for estimating the prevalence rate of members of a sampled population who fall in a rare subgroup. The paper examined the increase in precision and the increase in the number of subgroup members discovered in the samples, when shifting from the simplified single frame to the simplified dual frame design. This analysis was repeated for three differential per-unit costs. It showed improvements in each scenario, showing clear statistical benefits from the dual frame design.

This analysis has some important limitations, with the need for further research.
One limitation is that the effects of stratification were ignored on both frames in these scenarios. There may be limited effect on the outcomes regarding precision, if both approaches used the same stratification scheme. However, this is unclear at this point.

Second, if a dual frame approach were used in practice, it might ultimately be applied in the final stage of a multi-stage cluster sample. This would mean that the benefit to precision would be limited to the contribution to the variance of P only from the final stage of selection. Based on some initial thinking, the authors suspect that the effect of cluster sampling will be to dampen the relative reductions in the standard errors shown in this simplified comparison.

Some specific suggestions for subsequent research are to: (i) base comparisons between approaches on more realistic sample design and cost assumptions; and (ii) examine and find solutions through empirical research to all important operational barriers in working with administrative sources in the dual frame approach, with the cost and statistical implications of these barriers considered in evaluating the relative plausibility of the two design approaches.

5. Acknowledgements

The authors wish to acknowledge the contribution of the members of the Panel on Measuring Rape and Sexual Assault in Bureau of Justice Statistics Household Surveys, National Research Council.

6. References

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section.* American Statistical Association, 203-206.

Lohr, S. (2011). Alternative survey sample designs: sampling with multiple overlapping frames. *Survey Methodology, 37(2):194-213.*

National Research Council. (2013). *Estimating the Incidence of Rape and Sexual Assault.* Panel on Measuring Rape and Sexual Assault in Bureau of Justice Statistics Household Surveys, William D. Kalsbeek, Candace Kruttschnitt, and Carol C. House, Editors. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.