Published in final edited form as:
Stat Med. 2011 February 28; 30(5): 560–568. doi:10.1002/sim.3920.

Health Indicators: Eliminating bias from convenience sampling estimators

Bethany L. Hedt[†] and **Marcello Pagano**
Department of Biostatistics; Harvard School of Public Health; Boston, MA; USA

Abstract

Public health practitioners are often called upon to make inference about a health indicator for a population at large when the sole available information are data gathered from a convenience sample, such as data gathered on visitors to a clinic. These data may be of the highest quality and quite extensive, but the biases inherent in a convenience sample preclude the legitimate use of powerful inferential tools that are usually associated with a random sample. In general, we know nothing about those who do not visit the clinic beyond the fact that they do not visit the clinic. An alternative is to take a random sample of the population. However, we show that this solution would be wasteful if it excluded the use of available information. So, we present a simple annealing methodology that combines a relatively small, and presumably far less expensive, random sample with the convenience sample. This allows us to not only take advantage of powerful inferential tools, but also provides more accurate information than that available from just using data from the random sample alone.

Keywords

Prevalence; Surveillance; HIV; Bias; Sentinel Sites; Convenience Samples

1 Introduction

Monitoring and evaluating health programs is critical for ensuring successful and effective program implementation. This includes assessing the level, and changes in levels, of disease status in the population or measuring the coverage of program impacts. The challenge remains on obtaining these measures without depleting resources, an important consideration in all environments, but especially in countries with limited financial or human resources.

With the goal of conserving resources, these measures are often monitored in *convenient* sub-populations. These convenience samples are individuals easily accessible because they present in a central locale, such as a school or clinic, and the information on these individuals can be routinely collected as part of the program implementation. For example, disease status may be automatically measured and recorded for individuals who attend a clinic. Alternatively, a special activity may be designed in the convenient population, either by measuring all individuals or a sample of individuals. For example, sentinel surveillance activities monitor diseases in individuals at the site during a certain time period.

The benefit of monitoring diseases or program impacts in a convenience sample is that utilizing this population requires fewer resources than assessing the same measures in the

[†] Corresponding Author: 677 Huntington Ave; 4th Floor, Building 2; Department of Biostatistics; Boston, MA 02115; bhedt@hsph.harvard.edu; (w) 617-432-1056; (f) 617-432-5619.

general population. Because the convenient population is often geographically concentrated, the cost of collecting information or biological samples is lower than having to go from household to household. The actual activity can possibly be integrated into routine care, saving on hiring special staff for implementation. Even better, if the data is collected as part of routine services, the only added cost is the price of harvesting the data from site records.

However, the drawbacks of using convenience samples for program monitoring and evaluation is obvious. This population is rarely representative of the general population — there are often underlying, and unmeasured, attributes associated with membership of the convenient population and the measure of interest. Membership in the convenient population may indicate greater access to resources, better education or knowledge, social support, or even just geographic proximity to the clinic, any of which reasons can impact on the risk of disease or access to programs. Some claim that although an estimate based on a convenience sample is biased, that changes in disease prevalence or program impact over time in this convenient population reflects the change in the general population. However, the numerous conditions required for this assumption to hold are rarely met [1].

As a way of measuring disease or program impacts in the general population, full population surveys provide more representative samples for monitoring and evaluation. These surveys employ some statistical device — simple random sampling, cluster sampling, stratified sampling, or some combination of these — to randomly sample individuals from the population [2]. This gain in representation comes at added cost and complexity, so that these population surveys can only be implemented periodically providing little to no information in the period between surveys.

The conflict between estimates based on convenience samples and those based on full population samples is well documented in the literature when estimating HIV prevalence. In resource poor settings, population HIV prevalence estimates are often based on women attending antenatal care (ANC) clinics. A number of clinics are selected for inclusion, and residual samples of blood from women attending a clinic for the first ANC visit are tested for HIV. The estimates from this convenient population are then *extrapolated* to the general population. The quality of the estimates from these sentinel populations are mixed. In some cases, the prevalence in the ANC population closely approximates the general population. However, there is also evidence that the ANC prevalences can over- or under-estimate the prevalence in the general population, either overall or within age groups [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. Even when restricting inferences to all women, or even pregnant women, the ANC surveillance estimate has many potential biases [6, 9, 10, 14, 15, 16]. First, in many countries, not all pregnant women seek antenatal care in a clinic. Even with high coverage, the estimate is biased if the attendance at the clinic is associated with HIV status. For example, pregnancy, and therefore eligibility for ANC services, is an indication of recent unprotected sexual activity, which is clearly a predictor of infection. There is also evidence that HIV infected women have lower fertility, suggesting that women attending the clinic, especially in older age categories, have a lower HIV prevalence than the general population. Given the changing dynamics of HIV and ANC sites, it is unlikely that even changes in ANC sentinel surveillance estimates match the changes in HIV prevalence in the general population [9, 12, 16, 17, 18]. In light of these limitations, many recommend full population surveys as the best method to estimate HIV prevalence. However, full population surveys of any decent magnitude to provide accurate estimates are often too expensive and intensive to estimate HIV prevalence on a regular basis.

The methodology presented in this paper harmonizes the benefits of both convenience and population samples to support routine monitoring of health indicators. The *hybrid prevalence (*HP*) estimators* collect information from convenient populations, a low resource intensive activity, and anneals this data with small random samples to achieve unbiasedness. We show that these HP estimators are as or more efficient than population surveys that ignore the data from the convenience samples, and have the advantage of unbiasedness over convenience samples. For simplicity of exposition, we focus our derivations and comparisons on simple random samples, though the benefits are observed with any random sampling design. In the next section, we present the HP estimators followed by efficiency comparisons to illustrate their usefulness and accuracy.

2 Hybrid Prevalence Estimators

For every individual, we assume two measurements: one indicates the presence of a trait of interest and the second indicates membership in the convenient population, (X, Y). The variable X follows a Bernoulli distribution with probability of the trait of interest p, and the variable Y follows a Bernoulli distribution with parameter p_c — the probability that the individual attends the convenience site. For simplicity, we refer to the group of individuals that are not members of the convenient population as the *nonconvenient* population. Since there is often an association between the presence of a trait of interest and membership in the convenient population, we allow for the probability of the presence of the trait to be dependent on whether or not the individual accesses the site. Thus the probability the individual has the trait is p_s if the individual attends the site, and is p_s otherwise. Consequently,

$$p=p_c\,p_s+(1-p_c)\,p_{\bar{s}}. \quad (1)$$

This is the quantity we wish to estimate, the prevalence of the trait of interest in the general population, p. Written in this way, we can see that it is a function of three proportions, all of which can vary. In this paper, we investigate how differential and varying knowledge about all three of these quantities affects our hybrid prevalence estimators of p.

On occasion, the proportion with a particular trait in the convenient population, p_s, is used to estimate the proportion in the general population, p. We would then have the bias,

$$\begin{aligned} p-p_s &= p_c\,p_s+(1-p_c)\,p_{\bar{s}}-p_s \\ &= (1-p_c)\,(p_{\bar{s}}-p_s). \end{aligned}$$

As should be expected, the bias inherent in the estimator based solely on the convenient population decreases as the coverage of the convenience program gets bigger (p_c increases) or the two groups become more similar (p_s tends to p_s). This convenience sample only provides an unbiased estimator of p if the coverage of the convenient program is 100% (p_c = 1) or if the proportion of the population with the particular trait is the same in both the population accessing and those not accessing the convenience locale ($p_s = p_s$). These conditions are rather difficult to ascertain, but if violated, the estimator of the percentage of the overall population with a particular trait based solely on the convenient data is evidently biased. Further, even though the biases of a cross-sectional prevalence estimate may be recognized, sentinel systems are defended as a means of obtaining information on changes in prevalence over time, for example [7]. To see what happens to the above bias with time, attach a subscript of A to indicate an earlier time point, and a subscript of B to indicate a later time point. Then if we are interested in the change in prevalence with time, we have

$$p_A - p_B = \{p_{Ac}\, p_{As} + (1 - p_{Ac})\, p_{A\bar{s}}\} - \{p_{Bc}\, p_{Bs} + (1 - p_{Bc})\, p_{B\bar{s}}\}.$$

If we assume the access coverages remain the same ($p_{Ac} = p_{Bc}$), then we can simplify this expression:

$$p_A - p_B = p_c \{p_{As} - p_{Bs}\} + (1 - p_c)\{p_{A\bar{s}} - p_{B\bar{s}}\}.$$

So even with this simplifying assumption, the only way the change in prevalence in the general population is unbiasedly estimated with the change in prevalence in the convenient population is if the coverage of the convenient population is 100% (so $p_c = 1$) or if the changes are the same in the two groups ($p_{As} - p_{Bs} = p_{A\bar{s}} - p_{B\bar{s}}$).

We can obtain an unbiased estimate of the proportion of the population with a particular trait, such as HIV infection, by collecting information on the presence of the trait of interest, *X*, in a simple random sample (SRS) from the population at large. If additional information is also collected on membership in the convenient population, *Y*, so that each individual has two measurements, (X_i, Y_i), $i = 1, \ldots, N$, then the maximum likelihood estimator can be deconstructed into three pieces,

$$\begin{aligned}\widehat{p}_1 &= \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{N}\left\{\sum_{i=1}^{N} Y_i X_i + \sum_{i=1}^{N}(1 - Y_i)X_i\right\}\\ &= \frac{n_c}{N}\left\{\frac{1}{n_c}\sum_{i=1}^{n_c} X_i\right\} + \frac{n_{\bar{c}}}{N}\left\{\frac{1}{n_{\bar{c}}}\sum_{i=n_c+1}^{N} X_i\right\}\\ &\equiv \widehat{p}_c\, \widehat{p}_s + (1 - \widehat{p}_c)\widehat{p}_{\bar{s}}.\end{aligned}$$

Here, n_c is the number of people in the SRS that access the convenient locale (so $n_c = \sum_{i=1}^{N} Y_i$) and $n_{\bar{c}} = N - n_c$ is the number that do not. When compared to Equation 1, we see that this estimator simultaneously estimates all three parameters. The variance of the SRS estimator, written in terms of the three parameters, p_c, p_s, and $p_{\bar{s}}$, is

$$\begin{aligned}\mathscr{V}(\widehat{p}_1) &= \frac{1}{N}\{p_c\, p_s (1 - p_s) + (1 - p_c)\, p_{\bar{s}} (1 - p_{\bar{s}}) + p_c (1 - p_c)(p_s - p_{\bar{s}})^2\}\\ &= \frac{1}{N} p (1 - p).\end{aligned} \quad (2)$$

The SRS estimator provides an unbiased estimator of *p* in the general population. However, we contend that it is possible to unbiasedly estimate *p* while also increasing efficiency by incorporating information from a census or survey of the convenient population. For example, information routinely collected on the proportion of the convenient population with a particular trait, p_s, can be utilized to anneal the SRS population estimate. Additionally, from a regional census or program data, it may be known what proportion of the general population uses the services at the convenient location, p_c, thus making estimation of this parameter unnecessary. In the following subsections, we explore, in turn, the situations:

- when we know p_s, but do not know p_c or $p_{\bar{s}}$ ($\hat{p}_2$);
- when we know p_s and p_c, but do not know $p_{\bar{s}}$. Here we consider two situations: one, when we can identify and exclusively sample individuals in the nonconvenient

population $(\hat{p}_3)$ and two, when we have a mixed sample of individuals from the convenient population and nonconvenient population $(\hat{p}_4)$.

- when we know p_c, but only have sample estimates of p_s and p_s $(\hat{p}_5)$;

We present each of the proposed estimators and associated variances in this section, with proofs of unbiasedness and derivations of variances in the Supplemental Appendix. In Section 3, we demonstrate improvements in efficiency over the SRS estimator.

2.1 Complete Information from the Convenient Population

2.1.1 HP Estimators when p_c is Unknown—Information on traits of interest can be routinely collected on members of the convenient population as part of the normal operations of the site. If a prevalence estimate for the convenient population is based on a census, as in many sentinel surveillance systems or data from routine care, it would yield complete information about p_s. Thus, the first hybrid prevalence estimator we consider is one which supplements the SRS data with the information already obtained from the convenience population, by substituting for $\hat{p}_s$ the true proportion of the subpopulation with this particular trait, namely p_s. The resulting estimator is,

$$\begin{aligned}\hat{p}_2=&\tfrac{1}{N}\left\{\sum_{i=1}^{N}Y_i p_s+\sum_{i=1}^{N}(1-Y_i)X_i\right\}\\ \equiv&\ \hat{p}_c\,p_s+(1-\hat{p}_c)\,\hat{p}_{\bar{s}}.\end{aligned}$$

This estimator is also unbiased, with variance

$$\mathscr{V}(\hat{p}_2)=\frac{1}{N}\{(1-p_c)\,p_{\bar{s}}\,(1-p_{\bar{s}})+p_c\,(1-p_c)\,(p_s-p_{\bar{s}})^2\}. \quad (3)$$

2.1.2 HP Estimators when p_c is Known—In some situations, the coverage of services from the convenient population may be well established either through records or a previous census. In this case, it is unnecessary to re-estimate the proportion of the general population accessing the convenient locale, p_c. However, to estimate the overall proportion of the population with a particular trait, we must estimate the prevalence of the trait of interest in the population not attending the convenient site and combine this with information from the convenient population. In some instances, we can identify and sample the nonconvenient population without sampling anyone from the convenient population. We can fix the size of the sample from the nonconvenient population, n_c. The resulting estimator is then

$$\hat{p}_3=p_c\,p_s+(1-p_c)\overline{p}_{\bar{s}}$$

with

$$\overline{p}_{\bar{s}}=\frac{1}{n_{\bar{c}}}\sum_{i=1}^{n_{\bar{c}}}X_i.$$

We use the notation $p_{\bar{s}}$ since this estimator comes from a sample that *only* includes individuals from the nonconvenient population. It follows that this estimator is unbiased, with a variance of

$$\mathscr{V}(\widehat{p}_3)=(1-p_c)^2\, p_{\overline{s}}\, (1-p_{\overline{s}})/n_{\overline{c}}. \quad (4)$$

When it is not possible to separately identify individuals in this nonconvenient population, we cannot sample from them exclusively without also sampling individuals from the convenient population as well. In this case, the random sample will include individuals from *both* the convenient and nonconvenient populations, although only information from the populations not attending the convenient locale are included in the estimator. We then introduce

$$\widehat{p}_4=p_c\, p_s+(1-p_c)\widehat{p_{\overline{s}}},$$

with $\widehat{p_{\overline{s}}}=(\sum_{i=1}^{N}(1-Y_i)\, X_i)/n_{\overline{c}}$. Here we denote the estimator of the prevalence in the nonconvenient population as $\widehat{p_{\bar{s}}}$ (instead of $\bar{p}_{\bar{s}}$) to emphasize that the sample includes individuals from both the convenient and nonconvenient population, and only data collected on the nonconvenient population members is used. The variance of this estimator is the same as $\widehat{p_3}$; however, the costs of obtaining the same sample size, $n_{\bar{c}}$, from the nonconvenient population are much greater. Based on the negative binomial, the expected sample size to identify n_c individuals from the nonconvenient population is $N^* = n_c/(1 - p_c)$. This additional cost is accounted for in the efficiency comparisons discussed in Section 3.

2.2 Incomplete Information from the Convenient Population

Suppose we know the coverage of the convenient population, p_c, but from the routine activity, we only have information on a sample of M individuals from the convenient population. We consider the situation when we can get an exclusive random sample of the nonconvenient population. We end up with two random samples: 1) the simple random sample routinely collected from the convenient population and 2) the random sample collected from the nonconvenient population, which results in the following estimators,

$$\overline{p}_s=\frac{1}{M}\sum_{j=1}^{M}X_j \quad \text{and} \quad \overline{p}_{\overline{s}}=\frac{1}{n_{\overline{c}}}\sum_{i=1}^{n_{\overline{c}}}X_i.$$

These two independent random samples can be aggregated using stratified sampling theory, weighting each estimator by the proportion of the population the estimator represents [2]. Thus, our hybrid prevalence estimator is,

$$\widehat{p}_5=p_c\overline{p}_s+(1-p_c)\overline{p}_{\overline{s}}.$$

Based on stratified sampling theory, it follows that this estimator is unbiased with variance,

$$\mathscr{V}(\widehat{p}_5)=\frac{1}{M}p_c^2p_s\,(1-p_s)+\frac{1}{n_{\overline{c}}}(1-p_c)^2p_{\overline{s}}(1-p_{\overline{s}}). \quad (5)$$

3 Efficiency of the Hybrid Prevalence Estimators

Section 2 outlines various situations where convenience samples can be annealed with an SRS to obtain an unbiased estimator. The question then is, can we gain efficiency by incorporating all available data? Since all estimators are unbiased, we compare the SRS estimator to the hybrid prevalence estimators, in turn, via the relative efficiency, or the ratio of the variance of the SRS estimator to the hybrid prevalence estimator, $e_{1*} = \mathcal{V}(\hat{p_1})/\ \mathcal{V}(\hat{p_*})$. The relative efficiency is always at least one, and thus the variance of the HP estimator in every case is bounded above by the variance of the SRS estimator which does not incorporate any additional information (see proofs in the supplemental appendix). In all situations presented, the relative efficiency is a function of all three parameters p_c, p_s and p_s.$^-$ For each graph, we present results for all possible values of $p_s \in [0, 1]$. For simplicity, we limit discussions to two convenient population coverage levels, 90% and 50%; and we present three possible prevalences of the trait of interest in the nonconvenient population, p_s $^-$ = 0.2, 0.5, and 0.7, thus achieving a broad perspective. Figure 1 shows the overall prevalence of the trait of interest for these conditions.

First we look at the estimator for the situations described in Section 2.1.1, when the coverage of the convenient population is unknown. Since p_s is known with complete certainty, the term $p_c\ p_s\ (1 - p_s)$ no longer contributes to the variance of the estimator. Combining Equations 2 and 3, we have

$$\mathscr{V}(\widehat{p}_1)-\mathscr{V}(\widehat{p}_2)=\frac{1}{N}p_c\,p_s(1-p_s)\geq 0 \quad (6)$$

so that, indeed, there is a reduction in the variance with the second estimator. As expected, this reduction gets smaller as N gets bigger, and it gets bigger the better the convenience sample represents the population (increasing p_c). The relative efficiency, calculated by taking the ratio of the HP estimator to the SRS estimator, follows from Equations 2, 3 and 6,

$$e_{21}=\frac{\mathscr{V}(\widehat{p}_2)}{\mathscr{V}(\widehat{p}_1)}=1-\frac{p_c\,p_s(1-p_s)}{p(1-p)}.$$

This makes explicit the loss of efficiency if one ignores the data from the convenience sample. Of course, we can define its reciprocal, e_{12}, to show the gain in efficiency in using $\hat{p_2}$ instead of $\hat{p_1}$. Since $e_{21} \leq 1$, it follows immediately that e_{12} is bounded below by 1. This latter quantity is depicted in Figure 2. One sees that the biggest gain in efficiency occurs when p_c is highest and, conditional on p_c, when p_s and p are far from the extremities.

We derive two estimators in Section 2.1.2 for the situation when both p_c and p_s are known, leaving the parameter p_s as the only parameter that requires estimation. For the first estimator, we assume that individuals not attending the convenient locale are easily identified and are exclusively sampled. If we set the sample exclusively from this population as the same as the SRS, $n_c = N$, then combining Equations 2 and 4, we have

$$\mathscr{V}(\widehat{p}_1)-\mathscr{V}(\widehat{p}_3)=\frac{1}{N}\left[p_c\,p_s(1-p_s)+(1-p_c)p_{\overline{s}}(1-p_{\overline{s}})p_c+p_c\,(1-p_c)\,(p_s-p_{\overline{s}})^2\right]\geq 0.$$

Again, the difference decreases as N and p_c increase, but the difference is nonnegative, indicating that the variance of the hybrid prevalence estimator is always less than the SRS estimator. Figure 3 graphs the relative efficiency of the SRS estimator to the hybrid prevalence estimator, $e_{13} = \{p(1-p)\}/\{(1-p_c)^2\, p_s\,(1-p_s)\}$, which is also greater than or equal to one.

As described in Section 2.1.2, if we cannot sample the nonconvenient population exclusively but sample until we have n_c nonconvenient individuals, the final estimator and variance are the same as for $\hat{p}_3$. However, we must sample $N^*= n_c/(1-p_c)$ observations on average. Therefore, we will penalize the variance of $\hat{p}_4$, $\mathcal{V}(\hat{p}_4)$, by a factor of $1/(1-p_c)$ to more accurately reflect the true costs of obtaining this sample. Assuming $n_c = N$, and applying the penalty to $\mathcal{V}(\hat{p}_4)$, we have

$$\mathscr{V}(\widehat{p}_1)-\mathscr{V}(\widehat{p}_4)/(1-p_c)=\frac{1}{N}\left[p_c\,p_s\,(1-p_s)+p_c\,(1-p_c)\,(p_s-p_{\overline{s}})^2\right]\geq 0.$$

The relative efficiency of the SRS estimator to the resultant hybrid prevalence estimator, again applying the penalty of $1/(1-p_c)$ to $\mathcal{V}(\hat{p}_4)$, is

$$e_{14}=(1-p_c)\,e_{13}=\frac{p(1-p)}{(1-p_c)\,p_{\overline{s}}(1-p_{\overline{s}})}.$$

For all N, this is greater than 1 (see supplemental appendix). However, the gain in efficiency, penalized for the true sample size, is decreased from e_{13} by a factor of $1-p_c$.

Finally, we compare the estimators described in Section 2.2, when p_c is known, and we have information from a sample of the convenient population. For $\hat{p}_5$, we assume that individuals not attending the convenient locale are identifiable and can be sampled exclusively, resulting in a hybrid prevalence estimator based on stratified sampling. Here,

$$\mathscr{V}(\widehat{p}_1)-\mathscr{V}(\widehat{p}_5)=p_c\,p_s(1-p_s)\,\left[\frac{1}{N}-\frac{p_c}{M}\right]+(1-p_c)\,p_{\overline{s}}(1-p_{\overline{s}})\,\left[\frac{1}{N}-\frac{1-p_c}{n_{\overline{c}}}\right]+\frac{1}{N}p_c\,(1-p_c)\,(p_s-p_{\overline{s}})^2.$$

If we fix the sample size in the nonconvenient population to be the same as the SRS, $n_c = N$, and if the sample in the convenient population is at least as big or bigger than the simple random sample ($M \geq N$), then this difference is nonnegative. The difference decreases as N increases. The efficiency, combining Equations 2 and 5 and again fixing $n_c = N$, is

$$e_{15}=\frac{p\,(1-p)}{(N/M)p_c^2 p_s(1-p_s)+(1-p_c)^2 p_{\overline{s}}(1-p_{\overline{s}})}.$$

Meeting the conditions above, since $\mathcal{V}(\hat{p}_1) - \mathcal{V}(\hat{p}_5) \geq 0$, it follows that $e_{15} \geq 1$. Figure 4 shows the efficiency of $\hat{p}_5$ for varying sample size ratios comparing the size of the population SRS to the convenient sample, *N*/*M*. Clearly, the more information that can be obtained readily from the convenient sample, the bigger the increase in efficiency of the HP estimator over the SRS estimator.

4 Conclusion

The hybrid prevalence estimators originated with our search for effective ways to collect information on health indicators. Utilizing data from convenient populations, either data collected as a routine part of health care or through special sentinel surveillance activities, requires fewer resources and less time than collecting data from full population surveys. However, the inferences arising from these convenience samples are usually biased, and likely do not reflect the disease status or program impact in the general population. The complexity and costs of implementing and analyzing full population surveys, which should provide unbiased information, prohibits obtaining accurate information on health indicators on a regular and frequent basis.

The hybrid prevalence estimators provide gains in efficiency when full population surveys are supplemented with data easily and routinely available from convenient populations. We base the estimators and comparisons in this paper on SRS estimators to simplify the discussion, but the advantages readily extend to more complex sampling designs. Ultimately, the benefit of improved efficiency immediately translates into decreased sample sizes making it feasible to increase frequency of data collection to support data driven program management. Updating and modifying programs from current and accurate information improves the efficacy of health programs, and if obtainable at a reasonable cost, this is especially important for resource poor settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is supported in part by the National Institutes of Health grants T32 AI007358 and R01 EB006195. The authors thank Victor DeGruttola, George Seage, and Mariel Finucane for their insightful comments.

References

1. Hedt, Bethany L. PhD thesis. Harvard School of Public Health; 2008. Novel Methods for Efficient Surveillance and Monitoring.
2. Kish, Leslie. Survey Sampling. Wiley; 1995.
3. Ghys PD, Brown T, Grassly NC, Garnett G, Stanecki KA, Stover J, Walker N. The UNAIDS Estimation and Projection Package: a software package to estimate and project national HIV epidemics. Sex Transm Infect. Aug; 2004 80(Suppl 1):i5–i9. [PubMed: 15249692]
4. Dzekedzeke, Kumbutso; Fylkesnes, Knut. Reducing uncertainties in global HIV prevalence estimates: the case of Zambia. BMC Public Health. 2006; 6:83. [PubMed: 16579863]
5. Desgrées du Loû A, Msellati P, La Ruche G, Welffens-Ekra C, Ramon R, Dabis F. Estimation of HIV-1 prevalence in the population of Abidjan by adjustment of the prevalence observed in antenatal centres. AIDS. Mar; 1999 13(4):526–527. [PubMed: 10197385]
6. Fabiani, Massimo; Fylkesnes, Knut; Nattabi, Barbara; Ayella, Emingtone O.; Declich, Silvia. Evaluating two adjustment methods to extrapolate HIV prevalence from pregnant women to the general female population in sub-Saharan Africa. AIDS. Feb; 2003 17(3):399–405. [PubMed: 12556694]

7. Saphonn, Vonthanak; Hor, Leng Bun; Penh Ly, Sun; Chhuon, Samrith; Saidel, Tobi; Detels, Roger. How well do antenatal clinic (ANC) attendees represent the general population? A comparison of HIV prevalence from ANC sentinel surveillance sites with a population-based survey of women aged 15–49 in Cambodia. Int J Epidemiol. Apr; 2002 31(2):449–455. [PubMed: 11980815]

8. Ties Boerma J, Ghys Peter D, Walker Neff. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. Lancet. Dec; 2003 362(9399):1929–1931. [PubMed: 14667753]

9. Gouws E, Mishra V, Fowler TB. Comparison of adult HIV prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalised epidemics: implications for calibrating surveillance data. Sex Transm Infect. Aug; 2008 84(Suppl 1):i17–i23. [PubMed: 18647861]

10. Montana LS, Mishra V, Hong R. Comparison of HIV prevalence estimates from antenatal care surveillance and population-based surveys in sub-Saharan Africa. Sex Transm Infect. Aug; 2008 84(Suppl 1):i78–i84. [PubMed: 18647871]

11. Rice, Brian D.; Btzing-Feigenbaum, Jrg; Hosegood, Victoria; Tanser, Frank; Hill, Caterina; Barnighausen, Till; Herbst, Kobus; Welz, Tanya; Newell, Marie-Louise. Population and antenatal-based HIV prevalence estimates in a high contracepting female population in rural South Africa. BMC Public Health. 2007; 7:160. [PubMed: 17640354]

12. Fylkesnes K, Ndhlovu Z, Kasumba K, Mubanga Musonda R, Sichone M. Studying dynamics of the HIV epidemic: population-based data compared with sentinel surveillance in Zambia. AIDS. Jul; 1998 12(10):1227–1234. [PubMed: 9677172]

13. Kwesigabo G, Killewo JZ, Urassa W, Mbena E, Mhalu F, Lugalla JL, Godoy C, Biberfeld G, Emmelin M, Wall S, Sandstrom A. Monitoring of HIV-1 infection prevalence and trends in the general population using pregnant women as a sentinel population: 9 years experience from the Kagera region of Tanzania. J Acquir Immune Defic Syndr. Apr; 2000 23(5):410–417. [PubMed: 10866234]

14. Zaba BW, Carpenter LM, Boerma JT, Gregson S, Nakiyingi J, Urassa M. Adjusting ante-natal clinic data for improved estimates of HIV prevalence among women in sub-Saharan Africa. AIDS. Dec; 2000 14(17):2741–2750. [PubMed: 11125893]

15. Garnett GP, Grassly NC, Boerma JT, Ghys PD. Maximising the global use of HIV surveillance data through the development and sharing of analytical tools. Sex Transm Infect. Aug; 2004 80(Suppl 1):i1–i4. [PubMed: 15249691]

16. Glynn JR, Buv A, Caral M, Musonda RM, Kahindo M, Macauley I, Tembo F, Zekeng L. and Study Group on Heterogeneity of HIV Epidemics in African Cities. Factors influencing the difference in HIV prevalence between antenatal clinic and general population in sub-Saharan Africa. AIDS. Sep; 2001 15(13):1717–1725. [PubMed: 11546948]

17. Fylkesnes K, Musonda RM, Sichone M, Ndhlovu Z, Tembo F, Monze M. Declining HIV prevalence and risk behaviours in Zambia: evidence from surveillance and population-based surveys. AIDS. May; 2001 15(7):907–916. [PubMed: 11399963]

18. Ghys PD, Kufa E, George MV. Measuring trends in prevalence and incidence of HIV infection in countries with generalised epidemics. Sex Transm Infect. Apr; 2006 82(Suppl 1):i52–i56. [PubMed: 16581761]

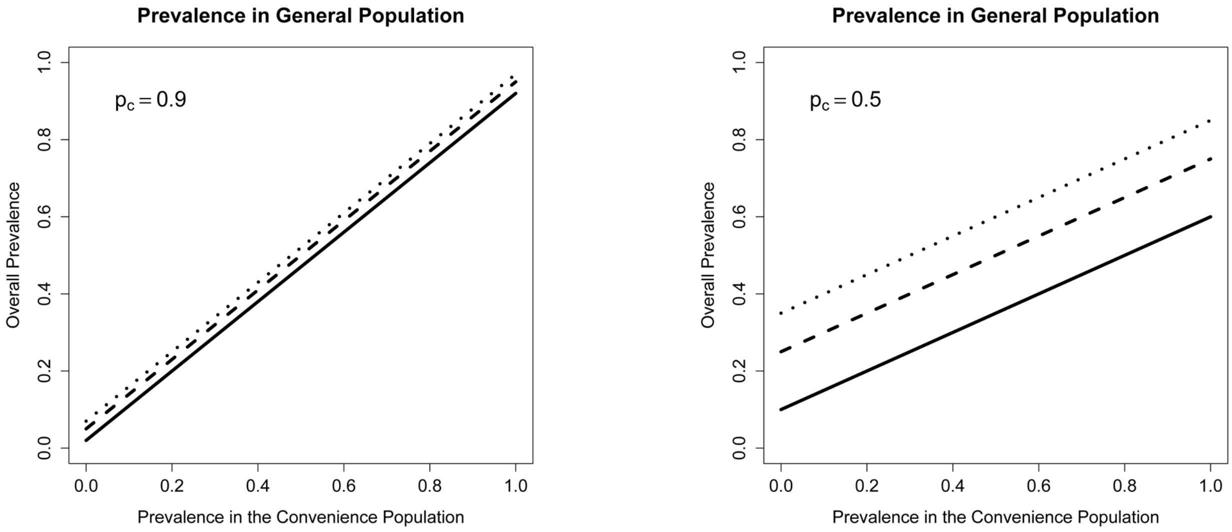

Figure 1.

Overall prevalence of the trait of interest in the general population. The plots on the left are for p_c = 0.9 and the plots on the right are for p_c = 0.5. The solid lines are for p_s = 0.2, the dashed lines are for p_s = 0.5, and the dotted lines are for p_s = 0.7.

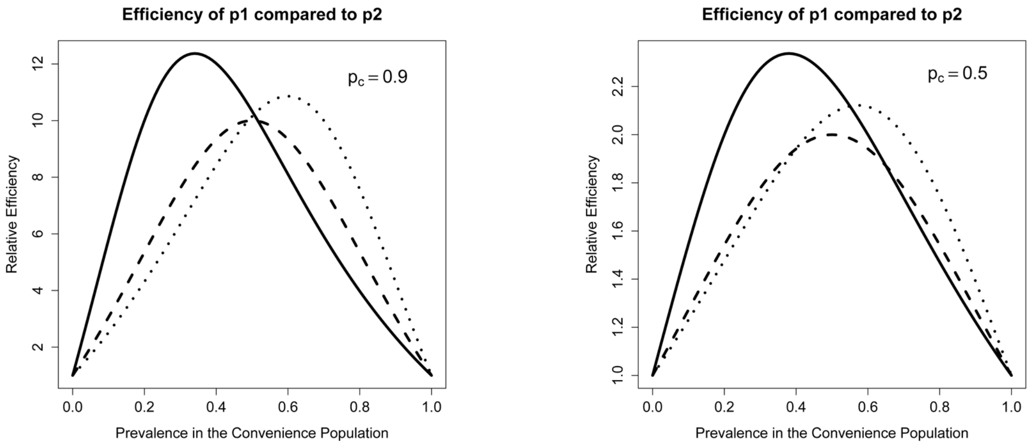

Figure 2.

Relative efficiency, e_{12}, of the estimator $\hat{p_1}$ relative to $\hat{p_2}$. The gain in using full information on the convenient population together with the simple random sample. The plots on the left are for $p_c = 0.9$ and the plots on the right are for $p_c = 0.5$. Note that the two vertical scales for relative efficiency are not the same. The solid lines are for $p_s = 0.2$, the dashed lines are for $p_s = 0.5$, and the dotted lines are for $p_s = 0.7$.

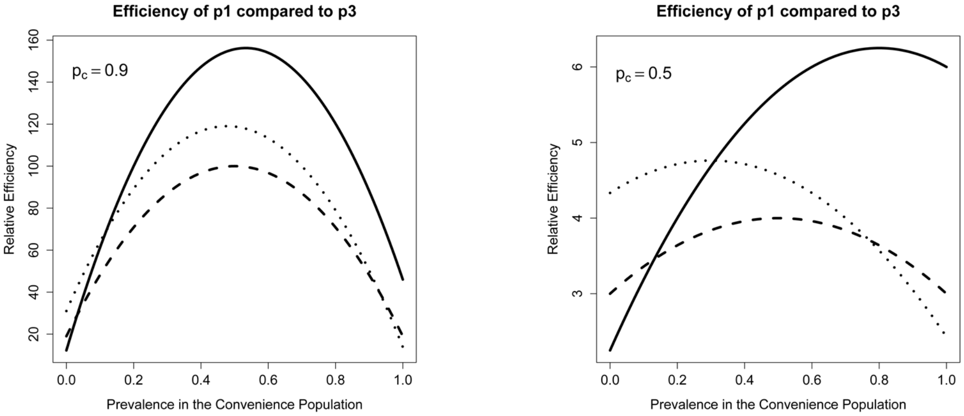

Figure 3.

Relative efficiency, e_{13}, of the estimator $\hat{p_1}$ relative to $\hat{p_3}$. The gain in using full information on the convenient population together with the simple random sample, when p_c is known. The plots on the left are for p_c = 0.9 and the plots on the right are for p_c = 0.5. Note that the two vertical scales for relative efficiency are not the same. The solid lines are for p_s = 0.2, the dashed lines are for p_s = 0.5, and the dotted lines are for p_s = 0.7.

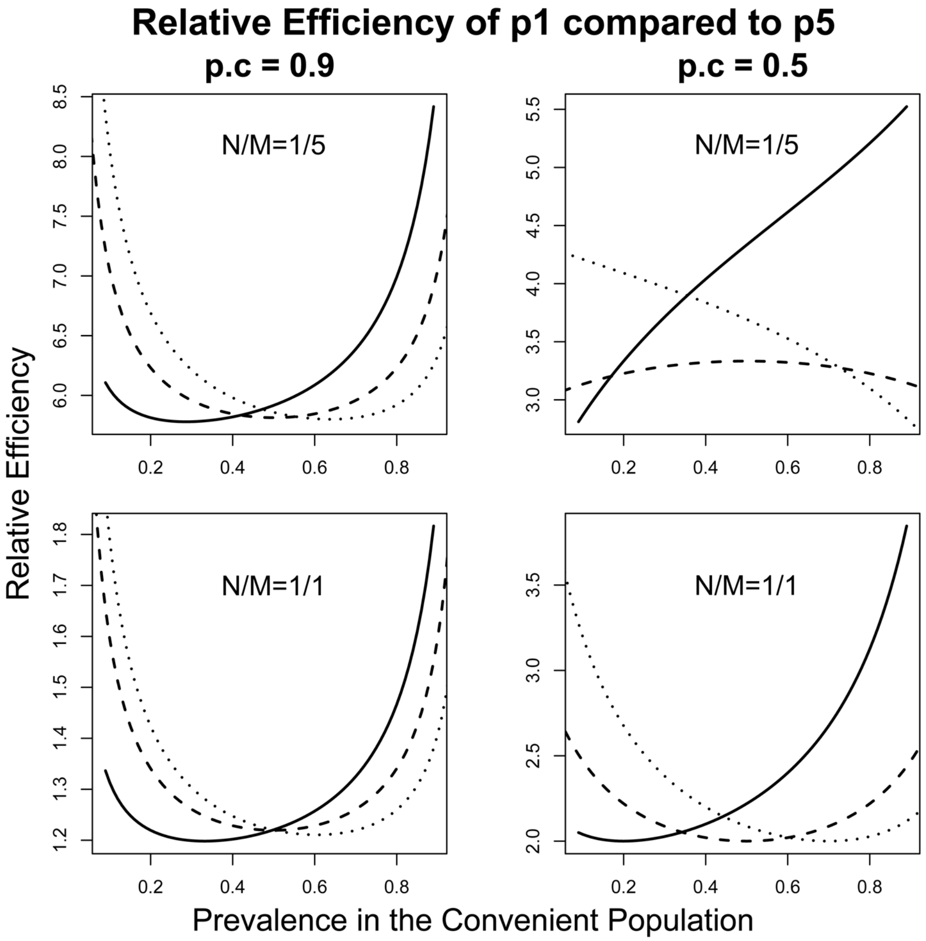

Figure 4.

Relative efficiency, e_{15}, of the estimator $\hat{p_1}$ relative to $\hat{p_5}$. The gain in efficiency using the convenience sample together with the simple random sample. Note that the two vertical scales for efficiency are not the same. The solid lines are for $p_s = 0.2$, the dashed lines are for $p_s = 0.5$, and the dotted lines are for $p_s = 0.7$.