RESEARCH REPORT SERIES
*(Statistics #2014-05)*

# A Small Sample Evaluation of Design-Adjusted Likelihoods Using Bernoulli Outcomes

Ryan Janicki
Donald Malec[1]

[1] National Center for Health Statistics

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: July 22, 2014

# A small sample evaluation of design-adjusted likelihoods using Bernoulli outcomes

Ryan Janicki[1] and Donald Malec[2]

[1]U.S. Census Bureau
[2]National Center for Health Statistics

November 25, 2013

**Abstract**

When there is a need to fit a parametric model to a finite population using data from a complex sample, the exact likelihood, which incorporates both the survey design and the probability model for the finite population, can be extremely complicated, or even intractable. Design-adjusted, approximate likelihoods, that is, modified likelihoods which incorporate the sampling design, are often used as an approximation to the exact likelihood. The design-adjusted, approximate likelihood can be used for either frequentist inference, or after specifying a prior distribution, Bayesian inference through the posterior distribution. The goal of this paper is to compare the design-adjusted, approximate likelihood to the exact likelihood, and to study the accuracy of this approximation. In this paper, two examples involving binary response data, the first using cluster sampling and the second stratified sampling, are presented in which the exact likelihood can be explicitly calculated. It is shown in these examples that even under extremely informative designs, the design-adjusted approximate likelihood closely matches the exact likelihood, with accuracy diminishing only for very small sample sizes or proportions close to the boundary. When lack of available resources and time constraints encourage the use of a design-adjusted, approximate likelihood, it is recommended that extreme designs, like the ones illustrated here, be used to help ascertain the scope of the error.

# 1 Introduction

Survey designs and subsequent randomization-based estimation procedures have evolved into a variety of complex methods that can be extremely cost-effective and provide estimates based on minimal assumptions. The design/randomization approach to sample selection/estimation offers a powerful and straightforward method for estimating a wide variety of finite population characteristics relatively quickly and efficiently. Besides providing the source for a wide variety of finite population estimates, including the model-assisted variety, data collected from a survey are also amenable to modeling. There is no inherent reason

why data from survey samples cannot be generalized beyond their finite population. However, modeling survey data presents many methodological challenges due to the possibility that complex sample selection can substantially change model relationships that would have been present had complex sampling not been used. Discussion on the difficulties of modeling complex survey designs can be found in Little (2004) and Gelman (2007).

Statistical model building often starts from the assumption that data are obtained via a simple random sample or, at most, are gathered from a designed experiment or a trial whose design-components are an integral part of the model. By contrast, using data arising from a survey for statistical modeling often introduces multi-layers of selection, weighting and clustering that are not of direct importance, and are possibly a detraction to the model (but may be related to the outcomes). The theoretical framework of modeling data from a survey and accounting for the fact that the design can influence the observations is discussed in Gelman et al. (1995), for example. The problem with modeling survey data is often more of a practical one, because accounting for the design through a model may necessitate extremely large models, with additional effort needed and with the possibility of over-parameterization due to lack of data available to account for all possible model and design factors and their interactions.

There are examples for which the exact likelihood can be constructed. Previous work includes Hartley and Rao, (1968), who derived the likelihood for $n_t$, the number of sampled units having one of a finite number of characteristics $y_t, t = 1, \ldots, T$, under simple random sampling without replacement, and showed that Bayesian inference could be performed by using a conjugate Dirichlet prior distribution. Similar work was done by Ericson (1969) and Hoadley (1969). Closely related is the Bayesian Bootstrap of Rubin (1981), which was extended by Aitken (2008). It should be noted that the likelihoods derived in these papers are the sampling likelihoods, and do not specify a distribution at the finite-population level.

If the goal is estimation of a parameter in a model for the finite population, or if the survey design is complex, finding the exact likelihood is more challenging. If the survey design is informative, in the sense that selection in the sample is related to the outcome variable, the finite population model could be different than the likelihood. One approach to reducing the impact of the survey design is to condition on a judicious choice of design variables, such as using sample design information as covariates in a regression model, enabling the use of standard, infinite population distributions and, hence, standard likelihoods methods for inference. A second approach is to use use random effects to account for clustering, such as in Malec and Sedransk (1985), Malec et al. (1997), and Sedransk (2008). Zheng and Little (2003, 2004, 2005) utilize spline models to account for sample selection proportional to size (PPS). A modeling approach to the selection probabilities, based on conditioning only on the sampled information, is provided by Pfeffermann, et al. (1998), in which some asymptotic justification for the approximate distribution was provided. Related work is Malec et al. (1999), which provided a semi-parametric, empirical Bayes approach to modeling the probability of selection using a marginal likelihood.

An alternative approach, which is the main focus of this paper, is to use design information to construct a function resembling a likelihood. When a parameter from a model for the finite population is of interest, approximations to the full-scale modeling of both the data collection and a population model are available, and the approximate likelihood can be constructed so that parameters from the population model will be included, so standard

methods for likelihood-based inference can be used. Examples of likelihood approximation for estimation of a model parameter include Asparouhov (2006), Raghunathan et al. (2007), Chen et al. (2011) and Joyce et al. (2013). Notable work includes the Bayesian pseudo-empirical-likelihood of Rao and Wu (2010) for estimation of a finite population parameter. They provide a design-weighted, possibly weight-calibrated modification to a empirical likelihood function. Although devised to cover a general parametric model based on the unique set of sampled outcomes, when applied to only binary responses (of sole focus, here) the pseudo-empirical-likelihood specific in equation (4) of Rao and Wu (2010) simplifies to the design-adjusted, approximate likelihood specified in equation (1).

This paper considers the accuracy a design-adjusted approximate likelihood for cases in which the design is essentially a nuisance parameter, and the underlying population model is of interest. In this paper, two examples are given where the exact likelihood can be computed explicitly under complex sampling schemes, and we examine how closely a likelihood approximation matches the exact likelihood. In these two examples the sampling scheme is informative, so the likelihood based on only the observed sampled data will be different than the population-level likelihood. In these examples, the objective is to estimate a proportion based on a small sample of binary response data, such as in small area or small domain estimation of characteristics or rates. These examples allow us to investigate when a design-adjusted likelihood can be a good approximation to the exact likelihood, and the extreme sampling designs used in these examples can be used to ascertain the error in using the approximate likelihood. This investigation may be particularly informative for Bayesian methods, in which posterior inference relies on the entire likelihood, so that the effects of misspecification of the likelihood can be severe.

This paper is organized as follows. Section 2 provides the background needed to specify the exact likelihood and the approximate, design-adjusted likelihood for the targeted example of binary data from a binomial superpopulation. Section 3 compares the exact and design-adjusted likelihood using extreme examples of a cluster sample. Section 4 provides a similar comparison for the case of a stratified sample. Appendix A summarizes the relationship between inference using a design-adjusted likelihood and inference using survey-weighted estimating equations, and provides context for the design-adjusted likelihood approach.

# 2 Background

## 2.1 The Exact Approach

In this paper we consider the setup where the finite population is conceptualized as consisting of independent random variables drawn from a parametric distribution (the "superpopulation" setup). This setup is required, for example, when the parameters of the population model are of primary interest. A sample from this finite population is then taken and observed according to a known sampling scheme. The likelihood based on the observed data from the survey must account for these two levels of randomization.

The steps to specifying the exact likelihood for sample data with a superpopulation model are straightforward and follow from the missing data paradigm of Rubin (1976). First, the complete likelihood must include both the model for the entire finite population as well as

the outcomes of those units selected in a sample (the sample design). Let $\underline{Y}$ denote a vector of outcomes (binary in this case) and $\underline{\delta}$ denote the vector of sample selection indicators (also binary) of the finite population. In general, a model for the joint distribution of $\underline{Y}$ and $\underline{\delta}$ is specified, conditional on unknown parameters. As this paper is primarily concerned with the modeling of binary survey data, the unknown parameter of interest is $p$, the superpopulaton probability that a population unit's outcome is a "success." The joint distribution of $\underline{Y}$ and $\underline{\delta}$ is denoted as $f\left(\underline{Y}, \underline{\delta} \mid p\right)$. Define $\underline{Y} = (y, \underline{Y}_{NS})$, where $y$ and $\underline{Y}_{NS}$ are the vectors of sampled and unsampled elements, respectively, of $\underline{Y}$, corresponding to $\underline{\delta}$. The distribution of the observed sample is obtained from the complete, joint distribution of sample indicators and the sampled and unsampled outcomes as follows:

$$\mathcal{L}\left(p\right) = f\left(y, \underline{\delta} \mid p\right) = \begin{cases} \sum_{\underline{Y}_{NS}} f\left(\underline{Y}, \underline{\delta} \mid p\right), & \text{if } \underline{Y} \text{ is discrete} \\ \int f\left(\underline{Y}, \underline{\delta} \mid p\right) d\underline{Y}_{NS}, & \text{if } \underline{Y} \text{ is continuous,} \end{cases}$$

which defines the likelihood.

As noted earlier, identifying a form for $f\left(\underline{Y}, \underline{\delta} \mid p\right)$ can be a major hurdle in this approach. In practice, a joint model is usually built using one of the two possible conditional specifications for the joint distribution:

$$f\left(y, \underline{\delta} \mid p\right) = f\left(\underline{\delta} \mid y, p\right) f\left(y \mid p\right)$$

or

$$f\left(y, \underline{\delta} \mid p\right) = f\left(y \mid \underline{\delta}, p\right) f\left(\underline{\delta} \mid p\right).$$

The first specification is known as the selection model in the missing data literature, while the second is the pattern-mixture model (Little and Rubin, 2002). The first approach is most useful if inference is based on the population model and the design parameters are irrelevant while the second is useful when the design parameters are included in the inferential model. While the second approach is usually easier to develop (such as by including random effects to account for clustering and fixed effects to account for stratification), the first approach is relevant when an underlying process, apart from the design, is of interest. Since the design-adjusted approximation we are investigating is an attempt at specifying the population model with the design parameters removed, we will only evaluate the first approach.

## 2.2 The Design Adjusted, Approximate Likelihood Approach

In a complex sampling design, elements that belong to a common cluster (or stratum) are typically more similar to each other than to elements of another cluster. Because of this, the number of sampled elements in a complex sample do not represent the distinct elements in a simple random sample with replacement of the same sample size, say $n$. Let $\hat{p}$ be the design-based estimator of $p$ under the current design and let $\hat{p}_{SRS}$ be an estimator of $p$ under a simple random sample design of the same sample size as the complex sample. The design effect for $\hat{p}$ (Kish and Frankel, 1974) is the ratio

$$Deff\left(\hat{p}\right) = \frac{\hat{V}_D\left(\hat{p}\right)}{\hat{V}_{SRS}\left(\hat{p}_{SRS}\right)},$$

where $\hat{V}_D(\hat{p})$ is a design-based estimate of the variance of $\hat{p}$, and $\hat{V}_{SRS}(\hat{p}_{SRS})$ can be approximated by $\hat{p}(1-\hat{p})/n$.

The effective sample size, $n'$, is defined as the ratio of the sample size to the design effect

$$n' = \frac{n}{Deff(\hat{p})} = \frac{\hat{p}(1-\hat{p})}{\hat{V}_D(\hat{p})},$$

and gives the sample size of a simple random sample with replacement that has the same precision (i.e., observed Fisher's information) as the realized sample under the complex design. The ratio $r = n'/n$, indicates the adjustment to the observed sample, $n$, needed to compensate for the sample selection. It will have a value of one if the design used was a simple random sample with replacement, and typically will be less than 1 for complex designs.

The design-adjusted, approximate likelihood for binary data can now be defined:

$$\mathcal{L}_{adj}(p) = p^{n'\hat{p}}(1-p)^{n'(1-\hat{p})}. \tag{1}$$

Note that under simple random sampling, $n' = n$, and (1) reduces to the standard binomial likelihood. The estimate that maximizes (1) is equal to $\hat{p}$, the design-based estimate, and the negative of the inverse of the second derivative of $\mathcal{L}_{adj}(p)$ evaluated at $p = \hat{p}$ is equal to $\hat{V}_D$, the design-based estimate of the design-based variance of $\hat{p}$ (see Appendix A for more detailed discussion). When only these two estimates are needed, the design-based estimates are preserved. However, inference based on the other features of the likelihood, such as tail area, has no design-based counterpart and the approximate and exact likelihoods may not provide similar results. In addition, (1) is not a likelihood because it is not the sampling distribution of $n'\hat{p}$. (If treated as a sampling distribution, (1) is a function of $p$ when summed over $n'\hat{p}$.)

This approach is in common use, and the main focus of this paper is understanding the strengths and weaknesses of using an approximation of the true likelihood. In this paper, two examples are given where the exact likelihood can be computed explicitly under complex sampling schemes, and we examine how closely (1) matches the exact likelihood. In these two examples the sampling scheme is informative, that is, $P(Y_i = y_i \mid i \in \mathcal{S}) \neq P(Y_i = y_i)$, where $\mathcal{S}$ is the set of indices included in the survey sample, so that the likelihood based on only the observed sampled data will be different than the population-level likelihood. In these examples, the objective is to estimate a proportion based on a small sample of binary response data, such as in small area or small domain estimation of characteristics or rates. These examples allow us to investigate when a design-adjusted likelihood can be a good approximation to the exact likelihood, and the extreme sampling designs used in the examples can be used to ascertain the error in using the approximate likelihood. This investigation may be particularly informative for Bayesian methods, in which posterior inference relies on the entire likelihood, so that the effects of misspecification of the likelihood can be severe.

# 3 Example 1: A Simple Random Cluster Sample

Consider the simple case when a finite population consists of an i.i.d. sample of Bernoulli($p$) outcomes. Further, suppose that a sample from the population is obtained via a one-stage simple random cluster sample without replacement, with equal size clusters, and that interest is in estimation of $p$ based on the observed sample. The following notation describes this situation:

- N: number of clusters,
- n: the number of clusters in sample,
- M: cluster size,
- Y: total number of responses coded as 1 out of $NM$,
- population model (before clustering): the units $j$ in cluster $i$ are independent, identically distributed Bernoulli random variables.

The population density function is then

$$f\left(Y \mid p, N, M\right) = \binom{NM}{Y} p^{Y} \left(1-p\right)^{NM-Y}. \tag{2}$$

If the entire population of $NM$ units were observed, the likelihood for the population could be obtained from (2) and inference would be straightforward. However, only the units in sampled clusters are observed. If the clustering mechanism were known, cluster membership could be modeled to remove the effects of the informative sampling design. However, when cluster membership is excluded from the population model, the design can be informative (as demonstrated in equations (4) and (5), below).

The effect of the most extreme clustering will be investigated so that the clusters formed are as homogenous as possible. By extreme clustering of all $NM$ units, we mean that clusters are composed completely of elements coded as 0, or completely of elements coded as 1, with the exception of one mixed cluster when $Y/M$ is not an integer.

Define $y$ to be the total number of observed 1's in the sample. Given $Y$, the sampling distribution is fixed. Let $\lfloor x \rfloor$ be the greatest integer less than or equal to $x$, and define $A = \lfloor Y/M \rfloor$ to be the number of clusters in the population consisting entirely of 1's. If $\mathrm{mod}(Y, M) = 0$, the number of sampled clusters consisting entirely of 1's, $a = y/M$, follows the hypergeometric distribution

$$f\left(a \mid A, Y, M, N\right) = \frac{\binom{A}{a}\binom{N-A}{n-a}}{\binom{N}{n}},$$

for $a \in \{\max\left(0, n+A-N\right), \ldots, \min\left(A, n\right)\}$. If $\mathrm{mod}(Y, M) \neq 0$, the sampling distribution consists of the number of sampled clusters, $a$, consisting entirely of 1's and whether of not the lone heterogeneous cluster is in sample, (i.e. $\Delta = 1$ or 0), so that $y = M * a + \Delta * \mathrm{mod}(Y, N)$.

For populations $Y$ of this type, the finite population sampling distribution follows a multiple hypergeometric distribution of the form

$$f\left(a, \Delta \mid A, Y, N, M\right) = \frac{\binom{A}{a}\binom{N-A-1}{n-a-\Delta}}{\binom{N}{n}}, \tag{3}$$

for $\Delta \in \{0, 1\}$ and $a \in \{\max(0, n + A - N), \ldots, \min(A, n - \Delta)\}$.

In general, the exact likelihood of $p$ can be obtained from the population distribution and the sampling distribution, and is given by

$$\mathcal{L}\left(p\right) = f\left(y \mid p, N, M, n\right) = \sum_{Y} f\left(y \mid Y, N, M, n\right) f\left(Y \mid p, N, M\right),$$

where $Y$ is summed over all population values consistent with the observed $y$ and $nM - y$.

For this example, when a sample of clusters is observed, the exact likelihood can be found by using the population density in (2) and the sampling density in (3). The form of this likelihood depends on whether the population count of "successes" is a multiple of $M$. If $\mathrm{mod}(y, M) = 0$,

$$\mathcal{L}\left(p\right) = \sum_{Y=y}^{(N-n)M+y} \frac{\binom{\left\lfloor \frac{Y}{M} \right\rfloor}{\frac{y}{M}}\binom{N-\left\lfloor \frac{Y}{M} \right\rfloor - I_{[mod(Y,M)\neq 0]}}{n-\frac{y}{M}}}{\binom{N}{n}} \times \binom{NM}{Y} p^{Y}\left(1-p\right)^{NM-Y}. \tag{4}$$

If $\mathrm{mod}(y, M) \neq 0$, then the one heterogeneous cluster has been sampled, and the remaining unsampled clusters must be completely homogeneous. The exact likelihood can then be found by summing over possible population clusters $A$ consisting entirely of $1's$:

$$\mathcal{L}\left(p\right) = \sum_{A=\left\lfloor \frac{y}{M} \right\rfloor}^{N-n+\left\lfloor \frac{y}{M} \right\rfloor - 1} \frac{\binom{A}{\left\lfloor \frac{y}{M} \right\rfloor}\binom{N-A-1}{n-\left\lfloor \frac{y}{M} \right\rfloor - 1}}{\binom{N}{n}} \binom{NM}{MA + \mathrm{mod}\left(y, M\right)} \times p^{MA+\mathrm{mod}(y,M)}\left(1-p\right)^{NM-MA-\mathrm{mod}(y,M)}. \tag{5}$$

## 3.1 Types Of Comparisons

The primary aim is to compare the design-adjusted, approximate likelihood, specified in (1) and described below in section 3.1.1, with the exact likelihood as derived in equations (4) and (5). In addition to this primary purpose, we include two other comparisons for contrast. These additional comparisons are: the binomial likelihood obtained if the sample design is ignored (in section 3.1.2) and a standardized design-adjusted, approximate likelihood (section 3.1.3); standardized so that, when summed over its support, it is not a function of $p$.

### 3.1.1 Design-adjusted Approximate Likelihood

For the combination of population distribution and sample design, the approximate design-adjusted approximate likelihood takes the form given in (1):

$$\mathcal{L}_{adj}(p) = p^{n'\hat{p}}(1-p)^{n'(1-\hat{p})}, \tag{6}$$

where

$$n' = \begin{cases} \hat{p}(1-\hat{p})/\hat{V}_D(\hat{p}) & \text{if } 0 < \hat{p} < 1 \\ n & \text{otherwise}^1 \end{cases} \tag{7}$$

and $\hat{V}_D(\hat{p})$ is a design-based estimate[2] of the variance of $\hat{p}$.

### 3.1.2 The Independent Likelihood: Ignoring The Sample Design

Ignoring the dependence caused by the clustering and subsequent cluster sample selection, a likelihood of the following form could be used:

$$\mathcal{L}_{ind}(p) = f_{ind}(y \mid p, M, n) = \binom{nM}{y} p^y (1-p)^{nM-y}. \tag{8}$$

### 3.1.3 Standardized Design-adjusted Likelihood

The design-adjusted likelihood is not a likelihood because its sum over the sample support will be a function of the parameter. One could devise a number of support spaces and calculate the proportionality constant so that that the likelihood represents an actual distribution. The following uses the original support space to evaluate the use of adjusting the adjusted likelihood so that it, once again, can be called a likelihood. In this case, the support of the sample is taken over the possible sample counts of $y$:

$$\mathcal{L}_{std}(p) = f_{std}(y \mid p, M, N, n) = \frac{\binom{nM}{y} p^{n'\hat{p}}(1-p)^{n'(1-\hat{p})}}{\sum_{y=0}^{nM}\binom{nM}{y} p^{n'\hat{p}}(1-p)^{n'(1-\hat{p})}}. \tag{9}$$

Although this is actually a likelihood (for something), it is a poor approximation to the exact likelihood as will be shown at the end of this section.

## 3.2 Results From Comparing The Exact And Approximate Likelihoods Of $p$

In this example, suppose that the finite population size is $NM = 1000$ and that equal-sized clusters of size $M = 10$ are formed. Figure 1 presents the likelihood and approximations when a sample of $n = 10$ clusters are selected. When 50 of the sampled binary observations

[1] If $\hat{p} = 0, 1$, the sample is completely homogeneous, so there is no empirical way to adjust for heterogeneity.

[2] $\hat{V}_D = \left(1 - \frac{n}{N}\right)\sum_{i \in S} \frac{(\hat{p}_i - \hat{p})^2}{n(n-1)}$

**Table 1:** Modal Values Corresponding to Figures 1a - 1c

| sampled successes | exact | independent | design-adjusted |
|---|---|---|---|
| 0.50 | 0.504 | 0.500 | 0.500 |
| 0.90 | 0.908 | 0.900 | 0.900 |
| 0.99 | 0.999 | 0.990 | 0.990 |

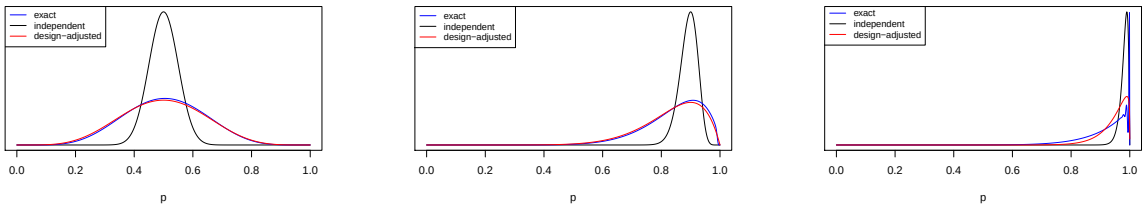
are "positive," the resulting exact likelihood and its approximations are displayed in Figure 1a. As can be seen, relative to assuming independence, the variance-adjusted likelihood is close to the exact likelihood. The likelihood assuming independence shares a similar peak but is too narrow. Similar conclusions can be drawn when 90/100 observations are successes, as seen in Figure 1b. However, the approximation appears less exact when nearly all observations are successes (Figure 1c).

Table 1 includes the maxima that correspond to Figures 1a - 1c confirming the visual observations. Figure 2 contains a close-up of the likelihood in Figure 1c revealing the multi-modal likelihood a little better: the reason for the multi-modal likelhood is the finite mixing of separate binomial likelihoods as specified in equations (4) and (5).

Figures 3 and 4 illustrate a few more cases of how the design may affect the approximation to the likelihood. In all of these examples, the total number of units sampled is 100, and results are presented when 50, 90 and 99 out of 100 "ones" are observed. Figure 3 essentially varies the cluster size and Figure 4 varies the total population of clusters. Note that in both figures, the plots of Figure 1 have been included to aid in comparison. As can be seen, the approximate likelihood deviates more from the exact likelihood when the cluster size is relatively large (Figure 3) or when the sampling fraction is large (Figure 4). In general, however, the exact and approximate likelihoods are similar to each other except in the extreme cases when $\hat{p}$ is near a boundary.

Since the exact distribution is based on the most extreme form of clustering and, hence, should exhibit the greatest difference from a simple random sample, it would be expected that the likelihood assuming i.i.d. Bernoulli random variables is too peaked. Other features of the likelihood appear to be different as well, so that caution must be taken if inference

**Figure 1:** Likelihood comparisons: N = 100, n = 10, M = 10




**(a)** 50/100 successes **(b)** 90/100 successes **(c)** 99/100 successes

about the tails of the distribution (or, sometimes, the modality) is of importance.

As seen in Figure 5, the standardized, approximate likelihood specified in Section 3.1.3 is a poor approximation to the exact likelihood (other, more skewed likelihoods that are not shown, are even more extreme). The main cause for the distortion is that standardization at the extremes of the likelihood require large factors. Figure 6 plots the adjustment factor (i.e. one divided by the denominator of equation (9)) on the log scale, and demonstrates the extreme weighting needed to standardize the adjusted likelihood for large and small values of $p$. Based on results like those shown in Figure 5, it is clear that this ad hoc standardization of the likelihood is problematic and will not be pursued further.

# 4 Example 2: Stratified Simple Random Sample

Stratification generally presents less of a problem to modeling since the strata are often included in the model. However, in cases like two-phase stratification, the entire population has not been stratified and either the stratification process needs to be modeled in order to draw inference about the unsampled population or an adjustment, such as the design-adjusted, approximate likelihood, could be used to try to remove the effects due to the design.

The following evaluates the possible errors in using a design-adjusted likelihood to counter the effects of stratification. As in the previous example, the underlying assumption is that the population is composed of i.i.d. Bernoulli random variables. As an illustration, only two strata will be considered. The following notation is needed:

- N: number of population units
- {A, B}: stratum identifiers
- $N_j$: number in stratum $j \in \{A, B\}$ (assumed known)
- $n_j$: sample size in stratum $j \in \{A, B\}$

**Figure 2:** Likelihood comparisons - magnified: 99/100 successes, N=100, n=10, M=10

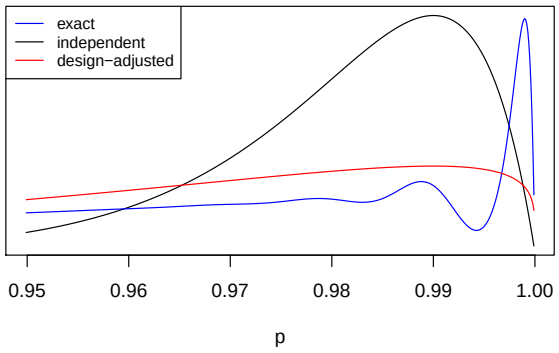
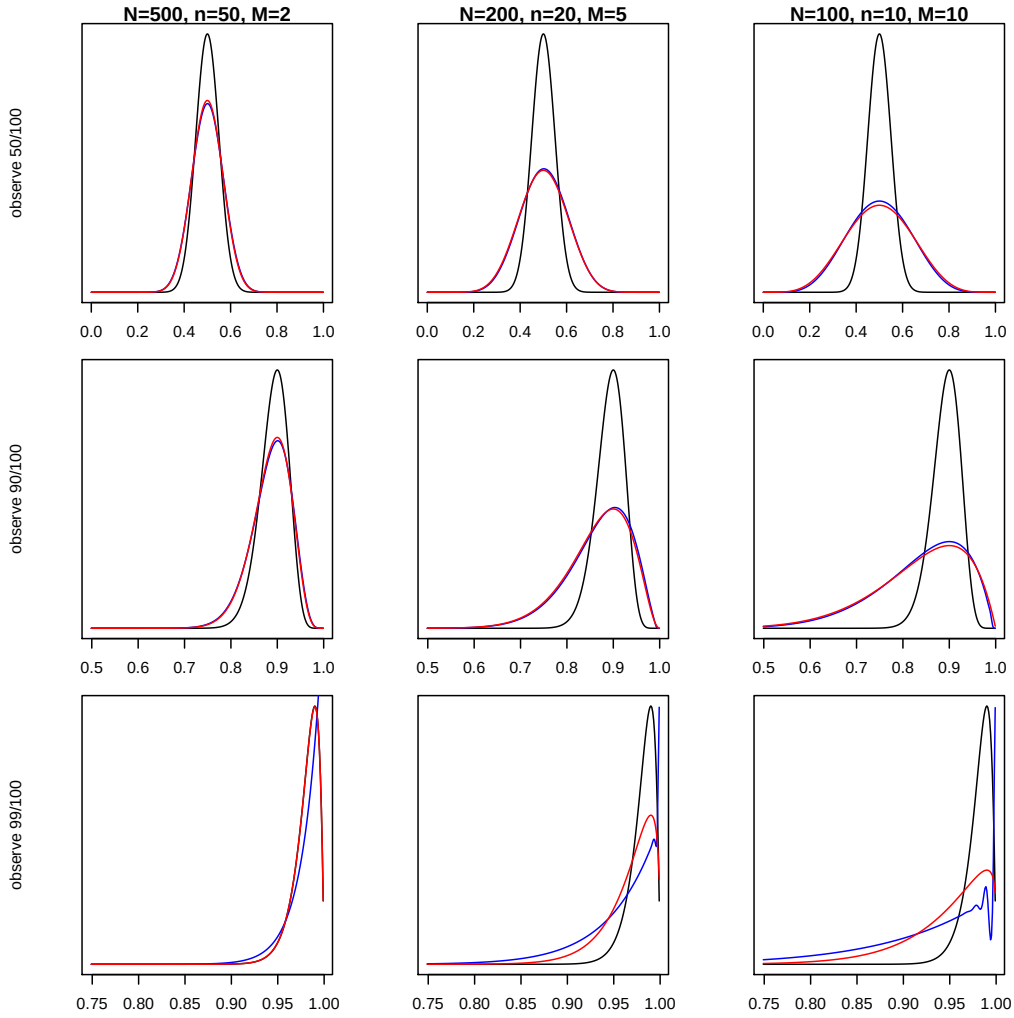


**Figure 3:** Comparisons when clusters sizes are 2, 5 and 10. —— Exact, —— Independent, —— Design-adjusted




- $m_{1j}$: number of sampled units which are "successes" (coded as "1") in stratum $j \in \{A, B\}$

As in Section 3, define $Y$ to be the total number of binary responses coded as a "1" out of $N$, so that,

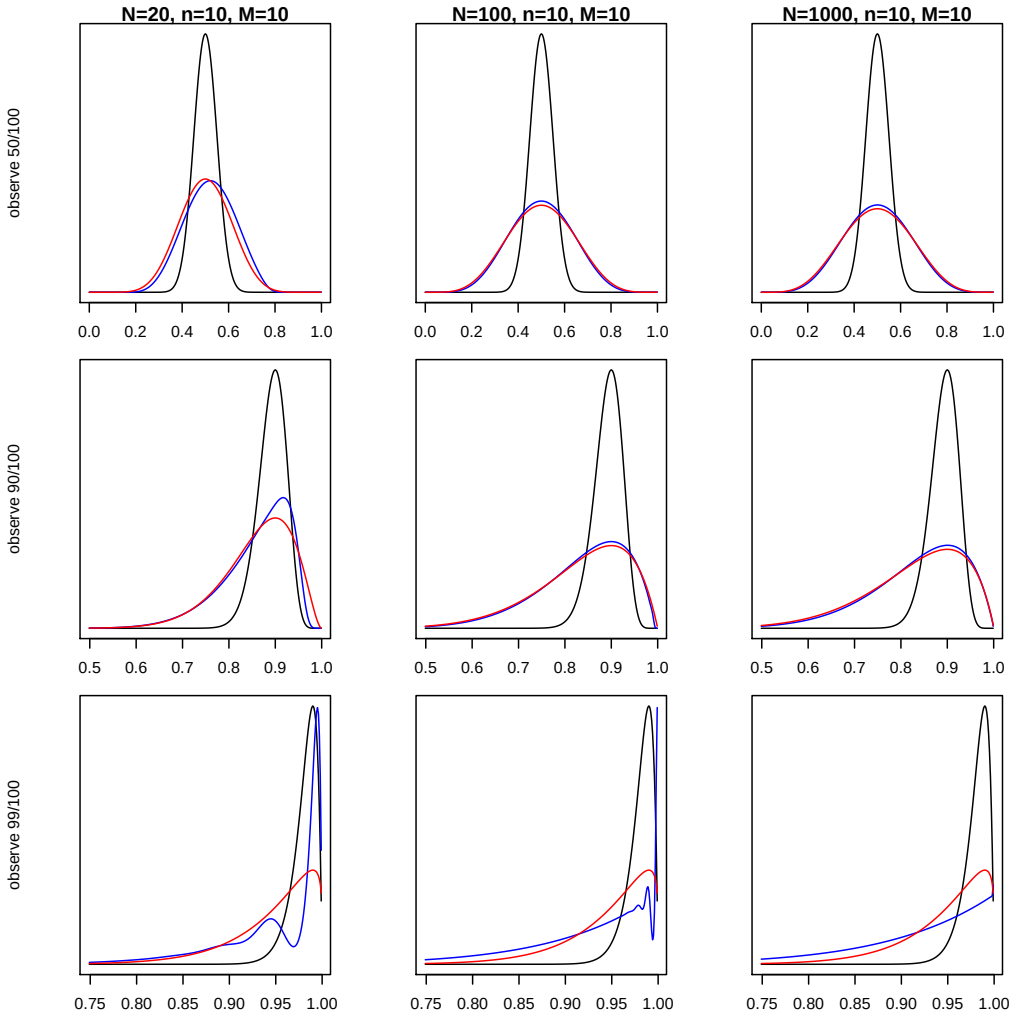
$$f\left(Y \mid p, N\right) = \binom{N}{Y} p^{Y} \left(1-p\right)^{N-Y}. \tag{10}$$

The allocation of the population units to strata will be modeled as independent Bernoulli outcomes dependent only on whether the unit is a success or not, so that

$$P\left(\text{unit in stratum A} \mid \text{ success}, \phi_1\right) = \phi_1,$$
$$P\left(\text{unit in stratum A} \mid \text{failure}, \phi_0\right) = \phi_0.$$

Define $Y_A$ to be the number of successes in stratum $A$ and $X_A$ to be the number of failures

**Figure 4:** Comparisons when total number of clusters are 20, 100 and 1000 —— Exact, —— Independent, —— Design-adjusted




in stratum $A$, so that

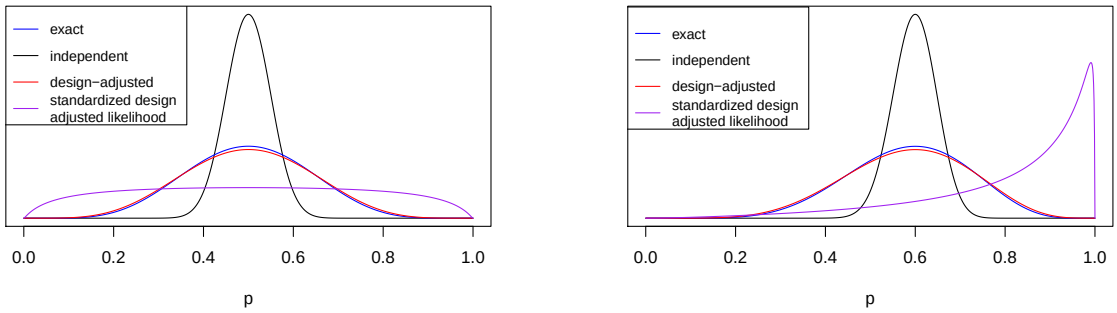
$$Y_A \mid Y, \phi_1 \sim Binomial\left(Y, \phi_1\right),$$
$$X_A \mid N, Y, \phi_0 \sim Binomial\left(N - Y, \phi_0\right).$$

The size of stratum $A$ is $N_A = Y_A + X_A$ and the size of stratum $B$ is $N_B = N - N_A$. Note that in this model formulation there is the possibility that only one stratum will be formed. The joint distribution of stratum membership is

$$f\left(Y_A, X_A \mid Y, N, \phi_1, \phi_0\right) = \binom{Y}{Y_A} \phi_1^{Y_A} \left(1 - \phi_1\right)^{Y - Y_A} \binom{N - Y}{X_A} \phi_0^{X_A} \left(1 - \phi_0\right)^{N - Y - X_A}. \quad (11)$$

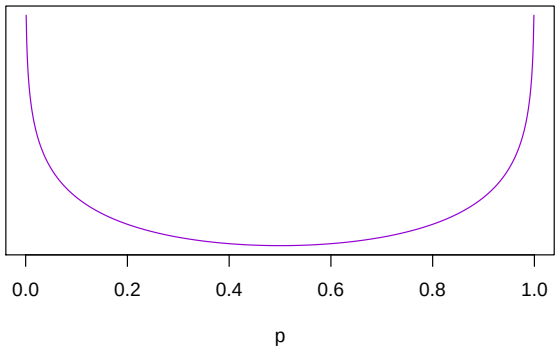
We assume that the stratum totals $N_A$ and $N_B$ are observable (but not $Y$, $Y_A$ or $Y_B$), and simple random samples of sizes $n_A < N_A$ and $n_B < N_B$ are selected without replacement

**Figure 5:** Likelihood comparisons with standardized design adjusted likelihood: N = 100, n = 10, M = 10




**(a)** 50/100 successes **(b)** 60/100 successes

**Figure 6:** Adjustment factor to standardized design adjusted likelihood (log scale)




from strata $A$ and $B$, respectively. The sampling distribution is then

$$f\left(m_{1A}, m_{1B} \mid Y, Y_A, N, N_A, n_A, n_B\right) = \frac{\binom{Y_A}{m_{1A}}\binom{N_A - Y_A}{n_A - m_{1A}}}{\binom{N_A}{n_A}} \times \frac{\binom{Y - Y_A}{m_{1B}}\binom{(N - N_A) - (Y - Y_A)}{n_B - m_{1B}}}{\binom{N_B}{n_B}}. \tag{12}$$

In this example, the goal is estimation of the parameter $p$, and the parameters $\phi_0$ and $\phi_1$ are considered nuisance parameters. By assuming a prior distribution $\pi\left(\phi_0, \phi_1\right)$, the exact likelihood with respect to $p$ can be found by summing over all unobserved random variables and integrating out the nuisance parameters $\phi_1$ and $\phi_0$. Here we use the uniform prior $\pi\left(\phi_0, \phi_1\right) = 1$. Let $\mathcal{S}_1 = \{m_{1A} + m_{1B}, \ldots, N - n + m_{1A} + m_{1B}\}$ and $\mathcal{S}\left(Y\right) = \{m_{1A}, \ldots, \min\left(N_A - n_A + m_{1A}, Y - m_{1B}\right)\}$, where $n = n_A + n_B$. Combining the densities

**Table 2:** Scenarios for comparing exact and approximate likelihoods

| scenario | stratum A size | sample allocation to stratum A | stratum successes proportional to stratum size |
|---|---|---|---|
| 1 | 0.5N | 0.5 proportional | no |
| 2 | 0.5N | proportional | no |
| 3 | 0.5N | proportional | yes |
| 4 | 0.75N | 0.5 proportional | yes |
| 5 | 0.75N | proportional | no |

(10), (11) and (12) gives an exact likelihood of

$$\begin{aligned}\mathcal{L}(p) = f(m_{1A}, m_{1B}, N_A \mid N, n_A, n_B, p) &= \sum_{Y \in \mathcal{S}_1} \sum_{Y_A \in \mathcal{S}(Y)} f(m_{1A}, m_{1B} \mid Y, Y_A, N, N_A, n_A, n_B, ) \\ &\times \iint f(Y_A, N_A \mid Y, N, \phi_1, \phi_0) \times \pi(\phi_0, \phi_1)\, d\phi_0\, d\phi_1 \times f(Y \mid N, p) \\ = \sum_{Y \in \mathcal{S}_1} \sum_{Y_A \in \mathcal{S}(Y)} &\frac{\binom{Y_A}{m_{1A}}\binom{N_A - Y_A}{n_A - m_{1A}}}{\binom{N_A}{n_A}} \times \frac{\binom{Y - Y_A}{m_{1B}}\binom{(N-N_A)-(Y-Y_A)}{n_B - m_{1B}}}{\binom{N_B}{n_B}} \times \frac{\binom{N}{Y} p^Y (1-p)^{N-Y}}{(N - Y + 1)(Y + 1)}.\end{aligned} \tag{13}$$

The term $1/(N - Y + 1)(Y + 1)$ comes from using a uniform prior on $\phi_0$ and $\phi_1$, making the appropriate transformation of (11), and integrating over the hyperparameters.

The total sample size, $n$, is fixed. As in any stratified sample design, the allocation of $n$ to strata is under the control of the survey design. Here, two types of sample allocation will be investigated: proportional allocation and oversampling by allocating sample to stratum $A$ one half the proportional amount (i.e., $n_A = 0.5nN_A/N$). Two possible total stratum sizes will also be investigated: equal size strata and strata size ratio of 3:1. Lastly, to complete the investigation, the effect of a disproportionate number of successes within strata is included as part of the evaluation as two possible situations; one, in which the number of stratum successes are proportional to the stratum size and the other where the number of stratum successes are as disproportionate as possible given the constraint due to the overall sample proportion. The five scenarios based on these criteria are shown in Table 2.

## 4.1 Types Of Comparisons

As stated earlier, our objective is to compare the design-adjusted, approximate likelihood with the corresponding exact likelihood. In this case, when the design is a stratified sample, the exact likelihood is specified by 13 and the design-adjusted, approximate likelihood is detailed in Section 4.1.1.

Although the comparison between the exact, marginal likelihood and the design-adjusted likelihood is our primary goal, as in section 3, we include comparisons with other approximate likelihoods for contrast. Here, we include a likelihood that would be appropriate if the sample

design was non-informative, as detailed in section 4.1.2, where neither the weights nor any other feature in the sample design needs to be accounted for in the likelihood. We also include a weighted likelihood, detailed in section 4.1.3 that uses the usual Horvitz-Thompson estimate of the population proportion but the only uses an adjusted likelihood that matches a simple random sample equal to the sample size.

### 4.1.1 Design-adjusted Likelihood

As described in Section 2.2, the design-adjusted likelihood takes the form

$$\mathcal{L}_{adj}(p) = p^{n'\hat{p}}(1-p)^{n'(1-\hat{p})}, \tag{14}$$

where

$$n' = \begin{cases} \hat{p}(1-\hat{p})/\hat{V}_D(\hat{p}) & \text{if } \hat{p} \neq 0, 1 \\ n & \text{otherwise}^3 \end{cases} \tag{15}$$

and $\hat{V}_D(\hat{p})$ is a design-based estimate[4] of the variance of $\hat{p}$.

### 4.1.2 Unweighted Likelihood

Disproportionate sampling in strata is reflected in the sampling weights. As a reference, however, the likelihood based on ignoring the design completely, (that is, ignoring the weights) will be used for comparison:

$$\mathcal{L}_U(p) = p^{m_{1A}+m_{1B}}(1-p)^{n-(m_{1A}+m_{1B})}. \tag{16}$$

### 4.1.3 Weighted Likelihood

An approximate likelihood often used in applications is one that uses the sample weights but adjusts the weighted counts to equal the sample total. Define the weighted likelihood to be

$$\mathcal{L}_W(p) = p^{n\hat{p}}(1-p)^{n(1-\hat{p})}, \tag{17}$$

where, $\hat{p} = (w_A m_{1A} + w_B m_{1B})/(w_A n_A + w_B n_B)$ and $w_A = N_A/n_A$, $w_B = N_B/n_B$.

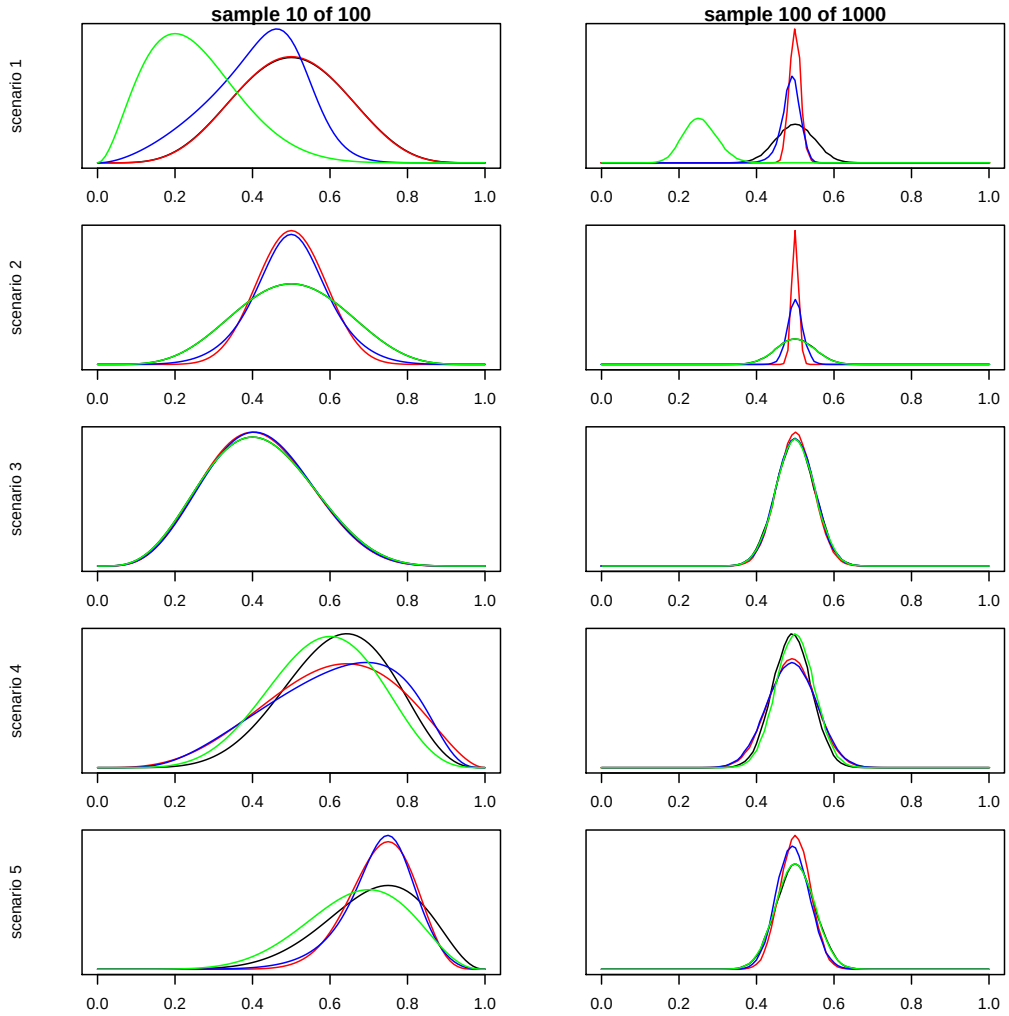
## 4.2 Comparisons Of The Exact and Approximate Likelihoods

This section presents comparisons based on the five scenarios defined in Table 2. Each of these five scenarios will be investigated for sample outcomes with a $\hat{p}$ near 0.5 and, also, $\hat{p}$ near 0.9; both for a very small sample size of ten and for a small sample size of 100. Figures 7 and 8 present results from the five scenarios based on these sample sizes. Figure 7 is for

[3] If $\hat{p} = 0$ or 1, the sample is completely homogeneous, so there is no empirical way to adjust for heterogeneity.

[4] $\hat{V}_D = \frac{1}{N^2}\sum_{k\in\{A,B\}}\left(\frac{N_k^2(N_k-n_k)}{N_k-1}\frac{\hat{p}_k(1-\hat{p}_k)}{n_k-1}\right)$, where $\hat{p}_k = m_{1k}/n_k$

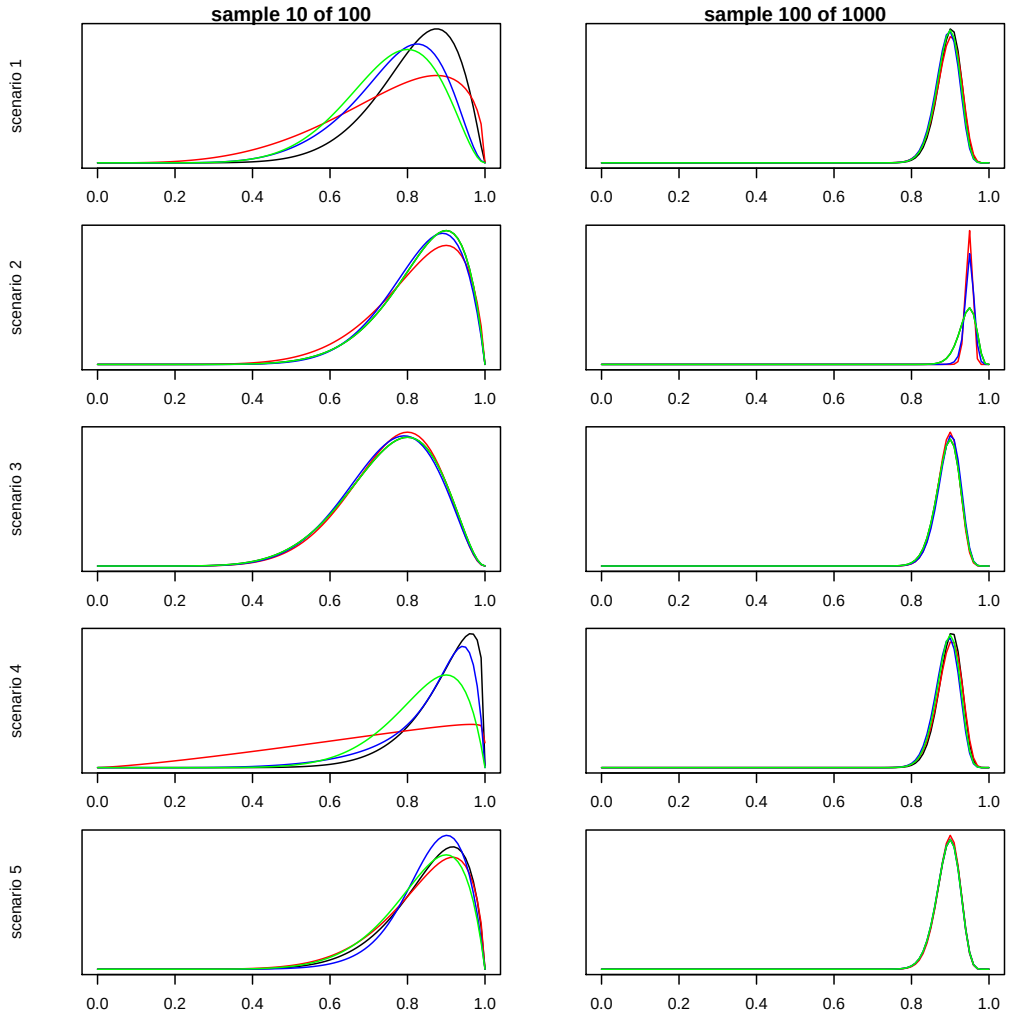
**Figure 7:** Likelihood Comparisons: $\hat{p}$ around 0.5, varied sample size. — Exact, — Weighted, — Design-adjusted, — Unweighted




outcomes in which the weighted estimate proportion is 0.5. Figure 8 presents results when the weighted sample proportion is 0.9.

Not surprisingly, there is little difference between any of the methods for scenario three (even the unweighted estimates) due to the fact that both the sample size and the sampled outcomes are proportional to stratum membership and stratum rate. In general, the non-proportional allocation illustrated in scenarios 1 and 4 produce the greatest deviation between the exact and either the weighted likelihood or the design-adjusted likelihood. In scenario 4, the design-adjusted likelihood follows the exact integrated likelihood fairly closely except when the sample size is very small. In summary, both using proportional allocation and drawing a relatively large sample improve the design-adjusted likelihood.

**Figure 8:** Likelihood Comparisons: $\hat{p} = 0.9$, varied sample size. —— Exact, —— Weighted, —— Design-adjusted, —— Unweighted




# 5 Summary

We have presented some simple examples showing that the design-adjusted approximate likelihood approach for inference is tenable even with some extremely informative designs. We have also identified some areas of concern. Given that the design-adjusted approach is relatively easy to implement, the question arises as to when it can be used in applications (especially in model-based estimation problems with small sample components). The examples presented here suggest a relatively quick way to evaluate the use of a design-adjusted likelihood for specific applications by devising an extreme design and evaluating the results against the corresponding exact likelihood. This should often be easier than trying to model the entire design. Even PPS designs can be quickly evaluated, in this manner, by grouping units with similar selection probabilities into strata and evaluating the approximate likelihood against an exact likelihood based on this approximate stratification.

Devising extreme designs should not be difficult once a sample is observed. For example,

given a cluster sample an extreme design could specify that all unsampled clusters were as homogeneous as possible (as in the examples here). A more realistic design would be to base the population distribution on the distribution of the sampled clusters.

The models in the two examples illustrated here are often used as the basic components of hierarchical models that are used in small area estimation and meta-analysis. Xie et al. (2006) show the dramatic effect of the distributional assumption on inference at the small area component level by evaluating t-distributions against a strict Normal distribution. In conclusion, the distributional assumptions used in these, often simple, components should be evaluated before being used to make inferences.

# Appendix

# A Similarities Between the Design-Adjusted Approximate Likelihood Approach and the use of Estimating Equations

Define unit $i$'s sample weight to be $w_i$ which is based on its selection probability and, additionally, sample-based non-response adjustment and post-stratification/coverage adjustment. The effective sample size, $n'$, can be used to approximate the likelihood by the product

$$A(p) = \prod_{i \in S} p^{n' w_i Y_i} (1-p)^{n' w_i (1-Y_i)}, \tag{18}$$

where $S$ is the set of sampled units.

Treating $A(p)$ as a likelihood allows to one use standard likelihood methods for inference. An estimator for $p$ can be found by maximizing $A(p)$, or equivalently, by finding a root of the derivative of the logarithm of (18), that is, a root of the equation

$$\sum_{i \in S} n' w_i \frac{Y_i - p}{p(1-p)} = 0, \tag{19}$$

to get the usual survey-weighted estimate for the population proportion

$$\hat{p} = \frac{\sum_{i \in S} w_i Y_i}{\sum_{i \in S} w_i}. \tag{20}$$

Note that (19) has a resemblance to a survey-weighted estimating equation (Binder, 1983),

$$\sum_{i \in S} w_i \frac{\partial \log f(y_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{21}$$

with the exception that, unlike the survey weights in equation (21), (19) contains the additional weighting factor, $n'$, that is also a function of the observed $Y_i$. In this particular application, however, the solution of (19) is identical to the solution of (21) when $f(y_i \mid \boldsymbol{\theta})$

is a Bernoulli distribution. Hence (19) involves an estimating function which can be identified with the survey-weighted estimating function in (21), which is optimal in the sense of Godambe and Thompson (1986). Unfortunately, in Bayesian inference, the optimality properties of Godambe and Thompson (1986) are not enough because the entire likelihood is used, not just the maximum and local curvature at the maximum. It may also be worth noting, that even the frequentist optimality properties in Godambe and Thompson (1986) are not applicable in general when design-based scaling factors (such as the value of $n'$ in equation (19)) vary across the sample.

There are several difficulties in attempting to use a likelihood-based analysis using survey data. The first is that the data from a survey are not independent realizations of a random variable, and when the survey design is informative, the form of the likelihood based on the observed survey data will be different from the population-level likelihood. The use of the design effect and the effective sample size is an attempt to correct for this, by approximating the number of independent observations that would give equivalent results to using the survey data.

Second, incorporating the scaling factor $n'$ and survey weights $w_i$ into (18) and treating $A(p)$ as a likelihood with independent factors can be misleading, in the sense that there is an implied reduction or inflation of the amount of information in the sample, depending on how the weights are scaled. Notice that the estimating function in (19) or the survey weights $w_i$ can be scaled by any constant factor, say $c$. If the survey weights are scaled by $c$, the resulting estimating equation

$$\sum_{i \in \mathcal{S}} cn' w_i \frac{Y_i - p}{p(1-p)} = 0$$

has the same point estimate $\hat{p}$ in (20) as a root. However, computing the Fisher information after scaling the weights, assuming $A(p)$ is the exact likelihood, gives $cn' \sum_{i \in \mathcal{S}} w_i / p(1-p)$ (compared to $n/p(1-p)$ in the i.i.d. setting). The scaling factor $c$ can be changed arbitrarily to give the illusion of increased information in the sample. For this reason, we assume the survey weights have been scaled so that $\sum_{i \in \mathcal{S}} w_i = 1$. Since the effective sample size $n'$ is less than, or equal to $n$, the information implied by the approximate likelihood $A(p)$ will not be greater than if the sample was a realization of i.i.d. random variables.

# Acknowledgements

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of either the U. S. Census Bureau or the National Center for Health Statistics, Centers for Disease Control and Prevention.

# References

Aitkin, M. (2008). "Applications of the Bayesian bootstrap in finite population inference." *J. Off. Statist.* **24**, 21 – 51.

Asparouhov, T. (2006). "General multi-level modeling with sampling weights." *Comm. Statist. Theory Methods* **35**, 439 – 460.

Binder, D. A. (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *Int. Statist. Rev.*, **51**, 279 – 292.

Chen, C. X., Lumley, T., and Wakefield, J. (2011). "The use of sample weights in Bayesian hierarchical models for small area estimation." Technical report. Available at `http://www.stat.washington.edu/research/reports/2011/tr583.pdf`.

Ericson, W. A. (1969). "Subjective Bayesian Models in sampling finite population." *J. Roy. Statist. Soc. Ser. B* **31**, 195 – 233.

Gelman, A. (2007). "Struggles with survey weighting and regression modeling." *Statist. Sci.* **22**, 153 – 164.

Gelman, A., Carlin, B. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*, Chapman and Hall, New York, Chapter 7.

Godambe, V. P. and Thompson, M. E. (1986). "Parameters of superpopulation and survey population: their relationship and estimation." *Int. Statist. Rev.* **54**, 127 – 138.

Hartley, H. O. and Rao, J. N. K. (1968). "A new estimation theory for sample surveys." *Biometrika* **55**, 547 – 557.

Hoadley, B. (1969). "The compound multinomial distribution and Bayesian analysis of categorical data from finite populations." *J. Amer. Statist. Assoc.* **64**, 216 – 229.

Joyce, P. M., Malec, D., Little, R. J. A., and Gilary, A. (2012). "Statistical modeling methodology for the Voting Rights Act Section 203 language assistance determinations." Technical report, U.S. Census Bureau, Center for Statistical Research and Methodology. Research Report Series. #2012-02.

Kish, L. and Frankel, M. R. (1974). "Inference from complex samples." *J. Roy. Statist. Soc. Ser. B* **36**, 1 – 37.

Korn, E. L. and Graubard, B. L. (1998). "Confidence intervals for proportions with small expected number of positive counts estimated from survey data." *Survey Methodol.* **24**, 193 – 201.

Little, R. J. (2004). "To model or not to model? Competing modes of inference for finite population sampling." *J. Amer. Statist. Assoc.* **99**, 546–556.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data.* second ed.

Wiley, Hoboken, New Jersey.

Malec, D., Davis, W. W., and Cao, X. (1999). "Model based small area estimates of overweight prevalence using sample selection adjustment." *Statist. Med.* **18**, 23, 3189-3200.

Malec, D. and Sedransk, J. (1985). "Bayesian inference for finite population parameters in multistage cluster sampling." *J. Amer. Statist. Assoc.* **80**, 897 – 902.

Malec, D., Sedransk, J., Moriarity, C. L., and LeClere, F. B. (1997). "Small area inference for binary variables in the national health interview survey." *J. Amer. Statist. Assoc.* **92**, 815 – 826.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998). "Parametric distributions of complex survey data under informative probability sampling." *Statist. Sinica* **8**, 1087 – 1114.

Rao, J. N. K. (2011). "Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal." *Statist. Sci.* **26**, 240 – 256.

Rao, J. N. K. and Wu, C. (2010). "Bayesian pseudo-empirical-likelihood intervals for complex surveys." *J. R. Statist. Soc. Ser. B* **72**, 533 – 544.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W., and Fueur, E. J. (2007). "Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening." *J. Amer. Statist. Assoc.* **102**, 474 – 486.

Rubin, D. B. (1976). "Inference and missing data." *Biometrika* **63**, 581 – 592.

Rubin, D. B. (1981). "The Bayesian bootstrap." *Ann. Statist.* **9**, 130 – 134.

Sedransk, J. (2008). "Assessing the value of Bayesian methods for inference about finite population parameters." *J. Off. Statist.* **24**, 495 – 506.

Stroud, T. W. F. (1994). "Bayesian analysis of binary survey data." *Canad. J. Statist.* **22**, 33 – 45.

Xie, D., Raghunathan, T. E. and Lepkowski, J. M. (2007). "Estimation of the proportion of overweight individuals in small areas - a robust extension of the Fay-Herriot model." *Statist. Med.* **6**, 2699–2715.

Zheng, H. and Little, R. (2003). "Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples." *J. Off. Statist.* **19**, 99–118.

Zheng, H. and Little, R. (2004). "Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples." *Survey Methodol.* **30**, 209 – 218.

Zheng, H. and Little, R. (2005). "Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model." *J. Off. Statist.* **21**, 1 – 20.