To Weight or Not to Weight:
Incorporating Sampling Effects into Model-Based Survey Analysis

Marianne Bertolet

Thesis Proposal
August 11, 2004

ABSTRACT

The two fundamental approaches used to analyze survey data, design- and model-based, differ in how they incorporate complexities of the sampling design, such as stratification, clustering and/or unequal probabilities of selection, into the survey analysis. The design-based analysis uses the sampling design as the sole source of the variability, which necessitates the use of survey sampling weights. Model-based analysis includes a model which generated the data in addition to the sampling design, which makes the role of the sampling weights questionable. The goal of this dissertation is to investigate the use of sampling weights in model-based analysis, and to that end, the proposed work has three major components. The first component investigates two current proposals for incorporating sampling weights to robustly estimate mixed-effects models in the presence of model mis-specification and/or informative sampling (Korn and Graubard, 2003; Pfeffermann et al., 1998), comparing their advantages and disadvantages in various situations. The second component addresses the role of sampling weights when the survey design was not formulated according to particular statistical model of substantive interest, and will consider models other than mixed-effects models to see how the work generalizes across model-based techniques. This component will consider using estimating equation methods (Binder and Patak, 1994) as a bridge between model-based and design-based analyses. Both of these components will be explored analytically, through simulation, and by considering examples involving current research questions of interest on large national surveys (such as the National Survey of Child and Adolescent Well-Being and/or the National Long Term Care Survey). As part of the thesis, guidelines regarding the use of sampling weights in model-based analysis will be proposed. The third component of this work will be to develop these recommendations as a set of publishable tutorials that serve to clarify these issues and methodologies for survey analysts.

1 Introduction

Large-scale statistical surveys seldom use simple random sampling. Two fundamental approaches (design- and model-based) exist regarding how to incorporate complexities such as stratification, clustering and/or unequal probabilities of selection into the survey analysis. Design-based analysis first originated around the early 1930's with Neyman (1934) and Fisher (1935). Model-based analysis began in the mid-1950's to mid-1960's with work by Godambe (1955, 1966) and Royall (1968). In the 1970's and 1980's, there were debates on the use and validity of the two types of analysis with Royall (1976), Hoem (1989) and Fienberg (1980, 1989) on the model-based side and Kalton (1968, 1989) and Hansen, Madow and Tepping (1983) on the design-based side, to name a few. Research from the 1990's continuing to today investigates when and how to combine the two approaches, for example see Little (1991), Pfeffermann (1993), Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998), Graubard and Korn(1996) and Korn and Graubard (2003). The survey world has been polarized regarding design and model-based analysis, and current research focuses on understanding how and when to combine them. The goal of this dissertation is to investigate the use of design-based sampling weights in model-based analysis, which involves three major components.

The first component investigates the use and estimation of mixed-effects models with complex survey sampling data. Research to date estimates mixed-effects model parameters (both fixed components and variance components) using both model- and design-based analysis (Pfeffermann et al., 1998 and Korn and Graubard, 2003). While these methodologies merge design- and model-based methodologies in one analysis, they lack guiding principles (Little, 2004). Current research focuses *how* to incorporate weights into mixed-effects models, and I will focus on *when it is appropriate* to incorporate weights into mixed-effects models. An example will be drawn from the National Survey of Child and Adolescent Well-Being (NSCAW), which analyzes fundamental questions about the outcomes of abused and neglected children who come into contact with the child welfare system.

The second component investigates the use of weights when the sampling design/data collection induces variance components not related to the researcher's question of interest, especially when it is addressed with a model other than a mixed-effects model in order to broaden the research on the use of weights beyond mixed-effects models. An example of this will be drawn from the National Long Term Care Survey (NLTCS), a national survey investigating health and behavioral factors associated with changes in chronic disability and mortality. A model which analyzes the cross categorical outcomes will be investigated.

The third component proposes better guidelines for the use of weights in model-based analysis. These guidelines will be explained and demonstrated through a series of tutorials designed to clarify the use of weights for survey analysts. These tutorials will include comparisons between design- and model-based analyses, and will examine different approaches to model-based analysis using both weighted and unweighted methods. The goal is to help analysts understand the issues for the use of sampling weights, and to aid them in making informed decisions regarding the weights in their own analyses.

The first component of this research regards the use of sampling weights specifically in mixed-effects models. The second component of this research is to investigate the use of weights when the sampling design and the model which generated the data do not have parallel variance components, specifically with the use of a model other than a mixed-effects model. The third component is to propose a set of guidelines and demonstrate them through a series of tutorials. The rest of this document proceeds as follows. Section 2 reviews background information on design-based versus model-based approaches and different optimality criteria used to judge estimators. Section 3 discusses the current research on the use of weights. Section 4 outlines the preliminary results obtained for this research. Section 5 proposes the work needed to complete this thesis.

2 Background

A goal of this research is to investigate the role of the sampling weights in design- and model-based analysis. This section describes both the traditional design-based approach and the model-based approaches. Key differentiating points between design- and model-based analyses include:

1. The population or estimand(s) about which we wish to make inference;
2. The source(s) of variability in the data we observe;
3. The role(s) of sampling weights.

The differences between design- and model-based analysis are summarized in Table 1. In addition to differences between design- and model-based analyses, many types of model-based analyses exist. This paper outlines three types, frequentist, alternate frequentist and Bayesian. A common example, when the sampling design is a stratified design, with simple random sampling within each strata (a non-informative sampling design), is used to compare the methods. This example is for the purpose of reviewing basic methodology. These assumptions are relaxed in Sections 3, 4 and 5, allowing for more complex designs (clustering and multi-stage sampling) and for informative sampling at any stage of the design.

2.1 Design-Based Analysis

The key differentiating points for design-based analysis are: 1) the population of interest is the specific finite population which was sampled, 2) the variability is induced by the sampling design and 3) the sampling weights are crucial for the analysis. See the solid arrows in Figure 1 . More formally, let $\mathcal{U}$ represent the finite population of N elements, and $\beta_{\mathcal{U}}$ be the target parameter to be estimated. This target parameter is a function of the finite population quantities, however census is not taken, so the parameter must be estimated from a sample of n elements. For a given sampling design, $p_{\mathcal{U}}(.)$, suppose there are t_n possible samples of size n from the population. Let $p_{\mathcal{U}}(i)$ be the probability of selecting the i^{th} sample and define $\hat{\beta}^{(i)}_{p_{\mathcal{U}}}$ to be the estimate of $\beta_{\mathcal{U}}$ from the i^{th} sample. Then $\{(\hat{\beta}^{(i)}_{p_{\mathcal{U}}}, p_{\mathcal{U}}(i))\}_{i=1}^{t_n}$ defines the randomization distribution (Lohr, 1999). The randomization distribution is also referred to as the sampling distribution.

As defined above, the estimates of the parameter from all the possible samples from the population are needed to know the randomization distribution. However, only one sample is taken from the population. To account for this, the sampling weights are used in the analysis. To define the sampling weights, let Z_{ki} be a random variable taking on the value 1 if, for example, the element i^{th} element in the k^{th} strata is included in the sample, and 0 otherwise. Note that the ki subscripts will vary according to the sampling design, for example, k may refer to clusters instead of strata, etc. Note that $E(Z_{ki}) = \pi_{ki}$, where π_{ki} is the probability that the ki^{th} element is included in the sample. The sampling weight is defined as $w_{ki} = \frac{1}{\pi_{ki}}$ (Korn and Graubard, 1999). An interpretation is that w_{ki} corresponds to the number of people which element ki represents in the population. These weights are important for two reasons; 1) in all but the simplest of sampling designs, design-based analyses without the weights have very large bias and 2) the design-based analysis gains information about the non-realized samples from the sampling weights and the sampling structure.

As an example of a design-based analysis, assume a stratified sampling design with K strata, and within each stratum there is simple random sampling (the sampling design is non-informative). The i^{th} element in the k^{th} stratum has an associated fixed quantity y_{ki} and random inclusion variable Z_{ki}. Let $\mathcal{U}_k$ represent the population in the k^{th} strata. Assume the population has N population elements, that stratum k has N_k population elements of which n_k are sampled. Suppose that target of interest is the mean of the finite population, denoted $\bar{y}^{(D)}$ where the superscript (D) denotes a design-based analysis. Let $\bar{y}^{(D)} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{U}_k} y_{ki}$.

A common estimator for the mean is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), $\hat{\bar{y}}_{HT}^{(D)}$ which is the weighted estimate of the sample elements. Thus,

$$\hat{\bar{y}}_{HT}^{(D)} \quad = \quad \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{U}_k} w_{ki} Z_{ki} y_{ki}.$$

Though the sum in the Horvitz-Thompson estimator appears to be over the entire population, the random inclusion indicator variables only allow elements in the sample to add to the sum term. This estimator is unbiased, because $E(Z_{ki}) = \frac{1}{w_{ki}}$. When considering the variance of the estimator, the covariance between the Z random variables is necessary. With this, the variance is

$$\text{Var}(\hat{\bar{y}}_{HT}^{(D)}) \quad = \quad \frac{1}{N^2} \sum_{k=1}^{K} N_k^2 \left(1 - \frac{n_k}{N_k}\right) \frac{\text{S}_k^2}{n_k}.$$

In the above equation, S_k^2 represents the sample variance of the sampled elements in the k^{th} stratum (Sarndal, Swensson and Wretman, 1992). The $1 - \frac{n_k}{N_k}$ term is often referred to as the finite population correction. Note that if a census is taken and $n_k = N_k$ then the variance becomes zero.

2.2 Model-Based Analysis

The key differentiating points for model-based analysis are: 1) the population of interest is either the specific finite population which was sampled or a superpopulation quantity, 2) the variability is induced a stochastic model assumed to generate variables associated with each element of the finite population and by the sampling design and 3) the use sampling weights is very controversial.

In model-based analysis, assume that for element ki there is an associated value Y_{ki} which was randomly generated by a model ξ. Consider this as a two-stage process, where the first stage generated the finite population of size N from the ξ model, and the second stage sampled the finite population. In this scenario, the estimation can be to a finite population parameter, or a superpopulation parameter. The finite population parameter estimation mimics the design-based analysis, while still assuming that the Y_{ki} value associated with element ki is random. A superpopulation estimand refers to a parameter (or function of parameters) from the generating stochastic model. See the dashed arrows in Figure 1.

The use of sampling weights in model-based analysis is controversial. Some researchers use sampling weights because they believe that they add robustness to the model-based analysis, reducing the dependence on the model assumptions (Kalton, 1989). Others estimate the parameters based on census data, and then include the weights as estimating the census estimates from the sample estimates (Pfeffermann et al. 1998, Korn and Graubard, 2003). Some researchers do not add sampling weights because adding the weights themselves imposes a model and assumptions which are not clearly explained. In addition, adding weights clouds interpretation and inflates variances (Hoem, 1989, Fienberg, 1989). As an interesting example, Hoem (1989) and Kalton (1989) debate over the use of weights in a Markov chain analysis. The controversy over the weights is whether or not to systematically put the sampling weights in as a way of including the randomization distribution, not whether they appear in the final estimate.

Three types of model-based analyses are presented in more detail in this section: frequentist, alternate frequentist and Bayesian. The frequentist is based on methods such as maximum likelihood estimation where the likelihood is based on both the sampling design and the generating model, see the dashed lines in Figure 1. The alternate frequentist approach uses frequentist methodology to obtain an estimate as if a census was taken, and then estimates that census estimate using design-based techniques, see Figure 2. The Bayesian approach takes the census likelihood, and integrates out the non-sampled values and makes inference based on posterior distributions. In some cases, the three model-based approaches produce the same estimates, however the assumptions behind them are quite different.

2.2.1 Frequentist Model-Based

As an example of frequentist model-based analysis, consider the superpopulation average, that is, the expected value of the average of any similarly constructed population created by the stochastic mechanism. Assume there is a stratified sampling design with K strata, and within each stratum there is simple random sampling (the sampling design is non-informative). The i^{th} element in the k^{th} stratum has an associated random quantity Y_{ki} and random inclusion variable Z_{ki}. Let $\mathcal{U}_k$ represent the population in the k^{th} strata, and S_k represent the sample from the k^{th} strata. Assume the population has N population elements, that stratum k has N_k population elements of which n_k are sampled. Let the data be generated as $Y_{ki} = \mu_k + \epsilon_{ki}$ where $\epsilon_{ki} \sim N(0, \sigma_k^2)$. The parameter of interest is $\bar{Y}^{(FM)}$, where the (FM) superscript denotes the frequentist model-based approach. Thus, $\bar{Y}^{(FM)} = \frac{1}{N}\sum_{k=1}^{K} N_k\mu_k$. To estimate $\bar{Y}^{(FM)}$, μ_k needs to be estimated. We know that the MLE for the mean of each stratum is the sample mean for the stratum. Inserting the MLE for μ, we get

$$\hat{\bar{Y}}^{(FM)} = \frac{1}{N}\sum_{k=1}^{K} N_k \sum_{i\in\mathcal{U}_k} \frac{1}{n_k} Z_{ki}Y_{ki} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in S_k} w_{ki}Y_{ki}.$$

The weights are included in this model-based analysis, not because they were inserted for the randomization distribution, but because they appeared due to numerical computations ($w_{ki} = N_k/n_k$ because of the simple random sampling). This estimator matches the design-based estimator, though it was derived under different assumptions. When taking the expected value of this estimator, both the distribution of the Y variables and the distribution of the Z variables are considered. As a result

$$E(\hat{\bar{Y}}^{(FM)}) = E_\xi E_p(\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{U}_k} w_{ki}Z_{ki}Y_{ki}) = E_\xi(\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{U}_k} Y_{ki}) = \frac{1}{N}\sum_{k=1}^{K} N_k\mu_k. \tag{1}$$

The variance is computed by conditioning on the finite population, Y_C,

$$\begin{aligned} \text{Var}(\hat{\bar{Y}}^{(FM)}) &= \text{Var}_\xi(E_p(\hat{\bar{Y}}^{(FM)}|Y_C)) + E_\xi(\text{Var}_p(\hat{\bar{Y}}^{(FM)}|Y_C)) \\ &= \frac{1}{N^2}\sum_{k=1}^{K} N_k\sigma_k^2 + \frac{1}{N^2}\sum_{k=1}^{K} N_k^2\left(1 - \frac{n_k}{N_k}\right)\frac{\sigma_k^2}{n_k}. \end{aligned} \tag{2}$$

The first term of the variance is the variance if a census were taken, and the second term of the variance is the added variance because a census was not taken. When a census is taken, the second term becomes zero.

2.2.2 Alternate Frequentist Model-Based Analysis

In the alternate frequentist model-based analysis, the estimate of the parameter is computed as if a census had been taken. This census estimate is estimated with the sample data using design-based techniques. This is one way to intentionally insert weights into a model-based analysis, see Figure 2.

As an example of an alternate frequentist model-based analysis, consider the superpopulation average, that is, the expected value of the average of any similarly constructed population created by the stochastic mechanism. Assume there is a stratified sampling design with K strata, and within each stratum there is simple random sampling (the sampling design is non-informative). The i^{th} element in the k^{th} stratum has an associated random quantity Y_{ki} and random inclusion variable Z_{ki}. Let $\mathcal{U}_k$ represent the population in the k^{th} strata, and S_k represent the sample from the k^{th} strata. Assume the population has N population elements, that stratum k has N_k population elements of which n_k are sampled. Let the data be generated as $Y_{ki} = \mu_k + \epsilon_{ki}$ where $\epsilon_{ki} \sim N(0, \sigma_k^2)$. The estimand is $\bar{Y}^{(AFM)}$, with the (AFM) superscript denoting the alternate frequentist model-based approach. Let $\bar{Y}^{(AFM)} = \frac{1}{N}\sum_{k=1}^{K} N_k\mu_k$. Next assume that we have the census data. We know that the MLE for μ_k is the mean of the data, in this case, the mean of the census values in stratum k. Substituting this in, we get $\hat{\bar{Y}}_C^{(AFM)} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{U}_k} Y_{ki}$. Next, estimate this using design-based methods. This was done in the design-based section above. Thus we get the alternate frequentist model-based estimator based on the sample data,

$$\hat{\bar{Y}}^{(AFM)} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in S_k} w_{ki}Y_{ki}.$$

The weights were explicitly added because of the randomization distribution. The expected value and variance of this estimate is evaluated with respect to both the p and ξ distributions. This AFM estimator is the is same as the FM estimator. Because the evaluation of the estimators is with respect to the same distributions, the expected value and variance are the same as in Equations 1 and 2.

$$\begin{aligned} E(\hat{\bar{Y}}^{(AFM)}) &= \frac{1}{N}\sum_{k=1}^{K} N_k\mu_k \\ \text{Var}(\hat{\bar{Y}}^{(AFM)}) &= \frac{1}{N^2}\sum_{k=1}^{K} N_k\sigma_k^2 + \frac{1}{N^2}N_k^2\left(1-\frac{n_k}{N_k}\right)\frac{\sigma_k^2}{n_k} \end{aligned}$$

In this example the frequentist model-based and alternate frequentist model-based estimators were the same. This only occurs in simple examples.

2.2.3 Bayesian Model-Based Analysis

The Bayesian model-based approaches differ from the frequentist model-based approaches, in that prior distributions are placed on the parameters of the model. All inference then is with respect to the posterior distributions of the parameters. Another difference is that the Bayesian model-based approach treats the non-sampled elements as missing values, and they are integrated out of the likelihood (Gelman, Carlin, Stern and Rubin, 1995). Links between Bayesian analysis and design-based analysis have been studied by Holt and Smith (1979) and Little (1991, 1993), to name a few.

As an example of Bayesian model-based analysis, consider the finite population average. This example differs from the others for two reasons: 1) the use of Bayesian methodologies instead of frequentist methodologies and 2) the use of a finite population estimand instead of a superpopulation estimand. The sampling design and generating stochastic mechanism are the same as the previous frequentist examples. Assume there is a stratified sampling design with K strata, and within each stratum there is simple random sampling (the sampling design is non-informative). The i^{th} element in the k^{th} stratum has an associated random quantity Y_{ki} and random inclusion variable Z_{ki}. Let $\mathcal{U}_k$ represent the population in the k^{th} strata, and S_k represent the sample from the k^{th} strata. Assume the population has N population elements, that stratum k has N_k population elements of which n_k are sampled. Let the data be generated as $Y_{ki} = \mu_k + \epsilon_{ki}$ where $\epsilon_{ki} \sim N(0, \sigma_k^2)$ are independent. The parameter of interest is $\bar{Y}^{(B)}$, where the (B) superscript denotes the Bayesian model-based approach. Let Y_C, Y_S and $Y_{C\setminus S}$ represent the census, sampled and non-sampled elements respectively. Then

$$\bar{Y}^{(B)} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{U}_k} Y_{ki} = \frac{n}{N}\bar{Y}_S + \frac{N-n}{N}\bar{Y}_{C\setminus S}.$$

In addition, assume that $\mu_k \sim N(\mu_{0k}, \sigma^2_{0k})$ where μ_{0k} and σ^2_{0k} are known and the μ_k's are independent. The census likelihood equation can be written as

$$p(Y_C, Z|\underline{\mu}) \;=\; \prod_{k=1}^{K}\prod_{i\in\mathcal{U}_k}\phi(Y_{ki}|\mu_k, \sigma^2_k)p(\{Z_{ki}=1\}),$$

where $\underline{\mu}$ refers to all the μ_k values and $\phi(Y_{ki}|\mu_k, \sigma^2_k)$ refers to the value of the normal density, with mean μ_k and variance σ^2_k, at Y_{ki}. This assumes that the sampling mechanism is independent of the data, or that the sampling is non-informative, which is the case with simple random sampling (SRS). In addition, the SRS in each strata is independent, so $p(\{Z_{ki}=1\}) = \frac{n_k}{N_k}$. The sample likelihood can be obtained by integrating non-sampled elements,

$$\begin{aligned}
p(Y_S, Z|\underline{\mu}) &= \int\prod_{k=1}^{K}\prod_{i\in\mathcal{U}_k}\frac{n_k}{N_k}Z_{ki}\phi(Y_{ki}|\mu_k,\sigma^2_k)\prod_{j\in\mathcal{U}_k}(1-Z_{ki})\phi(Y_{kj}|\mu_k,\sigma^2_k)dY_{C\setminus S} \\
&= \frac{n_k}{N_k}\prod_{k=1}^{K}\prod_{i\in\mathcal{U}_k}Z_{ki}\phi(Y_{ki}|\mu_k,\sigma^2_k).
\end{aligned}$$

Next, the posterior is calculated as

$$\begin{aligned}
p(\underline{\mu}|Y_S, Z) &\propto p(\underline{\mu})p(Y_S, Z|\underline{\mu}) \\
&\propto \left(\prod_{k=1}^{K}\phi(\mu_k|\mu_{0k},\sigma^2_{0k})\right)\left(\prod_{k=1}^{K}\prod_{i\in\mathcal{U}_k}Z_{ki}\phi(Y_{ki}|\mu_k,\sigma^2_k)\right) \\
&= \prod_{k=1}^{K}\left(\phi(\mu_k|\mu_{0k}\sigma^2_{0k})\prod_{i\in\mathcal{U}_k}Z_{ki}\phi(Y_{ki}|\mu_k,\sigma^2_k)\right) \\
&= \prod_{k=1}^{K}p(\mu_k|Y_S).
\end{aligned}$$

Let $\bar{Y}_{S_k}$ be the sample average of the Y values in the k^{th} stratum. Thus we see that the posterior of the μ_k's are independent and normally distributed, with the following mean and variance.

$$\begin{aligned}
E(\mu_k|Y_S, Z) &= \bar{Y}_{S_k}\left(\frac{n_k\sigma^2_{0k}}{n_k\sigma^2_{0k}+\sigma^2_k}\right) + \mu_{0k}\left(\frac{\sigma^2_k}{n_k\sigma^2_{0k}+\sigma^2_k}\right) \\
\text{Var}(\mu_k|Y_S, Z) &= \frac{\sigma^2_k\sigma^2_{0k}}{n_k\sigma^2_{0k}+\sigma^2_k}
\end{aligned}$$

Note that the posterior distribution of $\frac{1}{N}\Sigma_{k=1}^{K}N_k\mu_k$ would be used for Bayesian inference for the superpopulation parameter. To get the finite population estimate, the posterior predictive distribution is needed. We know that $p(\bar{Y}_{C\setminus S}|\underline{\mu})$ is a normal distribution with mean $\frac{1}{N-n}\Sigma_{k=1}^{K}(N_k - n_k)\mu_k$ and variance $\frac{1}{(N-n)^2}\Sigma_{k=1}^{K}(N_k - n_k)\sigma^2_k$. We showed above that $p(\underline{\mu}|Y_S)$ is normal. It can be shown that their multiplication, $p(\bar{Y}_{C\setminus S}, \underline{\mu}|Y_S)$ provided $Y_{C\setminus S}\perp Y_S|\underline{\mu}$, is normal. Thus, $\bar{Y}_{C\setminus S}|Y_S$ is a normal. By further conditioning on μ the mean and variance can be computed. To get this posterior, assume that $N_k \to \infty, n_k \to \infty$ and $\frac{N_k}{n_k}$ is constant. Pulling this together we get $p(\bar{Y}_{C\setminus S}|Y_S)$.

$$p(\bar{Y}_{C\setminus S}|Y_S) \sim \text{Normal} \quad \text{Mean} \approx \bar{Y}_S \quad \text{Variance} \approx \frac{1}{(N-n)^2}\sum_{k=1}^{K}N^2_k(1-\frac{n_k}{N_k})\frac{\sigma^2_k}{n_k}$$

Recall the goal is to estimate $\bar{Y}_C = \frac{n}{N}\bar{Y}_S + \frac{N-n}{N}\bar{Y}_{C\setminus S}$. Because $\bar{Y}_S$ is considered a constant to the posterior distribution, we can easily find the posterior distribution of $\bar{Y}_C$.

$$\bar{Y}_C \sim \text{Normal} \quad \text{Mean} \approx \bar{Y}_S \quad \text{Variance} \approx \frac{1}{N^2}\sum_{k=1}^{K} N_k^2(1-\frac{n_k}{N_k})\frac{\sigma_k^2}{n_k}$$

Note that the mean and variance of this posterior distribution match the mean and variance of the design based estimator.

2.3 Evaluation Criteria

Model-based survey sampling consists of two stages, first creating the finite population from the generating stochastic mechanism, and secondly obtaining a sample from the finite population. The criteria to evaluate the estimators becomes complicated in model-based analysis because of the need to take into account both the randomization distribution and the generating model distribution.

First consider the definition of these common criteria in the standard statistical theory. The bias of an estimator $\hat{\theta}$ for θ is $E(\hat{\theta}) - \theta$, and $\hat{\theta}$ is unbiased if the bias is zero. The definition of the variance of $\hat{\theta}$ is $E((\hat{\theta} - E(\hat{\theta}))^2)$ and the mean squared error (MSE) is defined as $E((\hat{\theta} - \theta)^2)$. The definitions of these quantities in survey sampling are very similar, as seen below. Now consider some asymptotic properties. Specifically, suppose that θ is estimated by $\hat{\theta}_n$, where $\hat{\theta}_n$ is a function of n independent and identically distributed random variables, say $\tau_1, \tau_2, ..., \tau_n$. Then $\hat{\theta}_n$ is asymptotically unbiased if $\lim_{n\to\infty}\hat{\theta}_n = \theta$, and it is consistent if, for any fixed ϵ, $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$. The problem with these criteria in survey sampling, is that the finite population size is fixed at N, and the sampling size $n < N$. Thus, another framework is needed to discuss these properties.

To properly define the asymptotic properties, consider the following notation (Hansen, et al. 1983 and Sarndal, Swensson and Wretman 1992). Let $y_1, y_2, y_3, ...$ be a sequence of variables of interest from an associated sequence of elements in the population, labeled $k = 1, 2, 3, ...$. A corresponding sequence of populations, $\mathcal{U}_1$,$\mathcal{U}_2$,$\mathcal{U}_3$, ..., can be defined, where $\mathcal{U}_v$ consists of the first N_v elements from the infinite sequence of elements. In this construction $\mathcal{U}_1 \subset \mathcal{U}_2 \subset \mathcal{U}_3, ...$ and $N_1 < N_2 < N_3 < ....$ Let $\theta_{\mathcal{U}_v}$ be the value of a parameter in population $\mathcal{U}_v$. Each population $\mathcal{U}_v$ has a sampling design $p_{\mathcal{U}_v}()$ which assigns probability $p_{\mathcal{U}_v}(s_v)$ to sample s_v.

Assume the sample sizes are fixed at n_v, where $n_1 < n_2 < n_3 < ....$ Let $\hat{\theta}_{p_{\mathcal{U}_v}}$ be an estimator of $\theta_{\mathcal{U}_v}$ based on the sampling design $p_{\mathcal{U}_v}$. Hansen, et al. (1983) and Sarndal, Swensson and Wretman (1992) define the following:

- p-bias and p-variance – An estimator $\hat{\theta}_{p_{\mathcal{U}_v}}$ is said to be *p-unbiased* for a parameter $\theta_{\mathcal{U}_v}$ in population $\mathcal{U}_v$ if $E_{p_{\mathcal{U}_v}}(\hat{\theta}_{p_{\mathcal{U}_v}}) = \sum_{s_v \in \mathcal{U}_v} \hat{\theta}_{p_{\mathcal{U}_v}}(s_v) p_{\mathcal{U}_v}(s_v) = \theta_{\mathcal{U}_v}$, where $s_v \in \mathcal{U}_v$ represents all possible samples in the finite population $\mathcal{U}_v$ and $\hat{\theta}_{p_{\mathcal{U}_v}}(s_v)$ is the estimator of $\theta_{\mathcal{U}_v}$ based on the sample s_v. Similarly, the *p-variance* of the estimator $\hat{\theta}_{p_{\mathcal{U}_v}}$ is defined as Var$(\hat{\theta}_{p_{\mathcal{U}_v}}) = E_{p_{\mathcal{U}_v}}(\hat{\theta}_{p_{\mathcal{U}_v}} - E_{p_{\mathcal{U}_v}}(\hat{\theta}_{p_{\mathcal{U}_v}}))^2$.

- p-MSE – The *p-mean squared error* for $\hat{\theta}_{p_{\mathcal{U}_v}}$ is defined as $E_{p_{\mathcal{U}_v}}(\hat{\theta}_{p_{\mathcal{U}_v}} - \theta_{\mathcal{U}_v})^2$

- p-asymptotically unbiased – An estimator $\hat{\theta}_{p_{\mathcal{U}_v}}$ is *p-asymptotically unbiased* for $\theta_{\mathcal{U}_v}$ if $\lim_{v\to\infty}[E_{p_{\mathcal{U}_v}}(\hat{\theta}_{p_{\mathcal{U}_v}}) - \theta_{\mathcal{U}_v}] = 0$

- p-consistent – An estimator $\hat{\theta}_{p_{\mathcal{U}_v}}$ is *p-consistent* for $\theta_{\mathcal{U}_v}$ if $\lim_{v\to\infty} \Pr(|\hat{\theta}_{p_{\mathcal{U}_v}} - \theta_{\mathcal{U}_v}| > \epsilon) = 0$

- p-consistent for a finite population – An estimator $\hat{\theta}_{p_{\mathcal{U}_v}}$ of $\theta_{\mathcal{U}_v}$ is *p-consistent for a finite population* under a given class of designs if the sample equals the population, i.e. $s_v = \mathcal{U}_v$ implies that $\hat{\theta}_{p_{\mathcal{U}_v}}(s_v) = \theta_{\mathcal{U}_v}$.

The definitions regarding the asymptotics are still somewhat nebulous because the limit process has not been fully specified. For example, in a cluster sample design, as the size of the population grows, the number of sampled clusters can grow while the number of people within each cluster remain constant, or the number of clusters can remain constant as the number of people within each cluster grow.

To evaluate the model-based estimators, both the p and the ξ distributions are important. There are many different ways they can be combined, to create many different criteria (Sarndal, 1978). For example: ξp unbiased ($E_\xi E_p(\hat{\theta}_p) = \theta$) and ξp variance ($\text{Var}_\xi(E_p(\hat{\theta}|Y_C)) + E_\xi(\text{Var}_p(\hat{\theta}|Y_C))$) as used in the previous section. These types of criteria will be used when comparing and evaluating the different estimators presented.

3 Current Approaches

With a basic understanding of the differences between design- and model-based analysis, along with how each uses the weights, current research on the three components of this dissertation will be reviewed.

3.1 Weights in Mixed-Effects Models

The first component of this research is to investigate the use of sampling weights, specifically in mixed-effects models. Before doing this, it is necessary to understand how to incorporate the weights in these models. This section reviews linear mixed-effects models and compares current weighting methods.

Linear mixed-effects (LME) models are used to analyze clustered data, for example for analyzing children in classrooms. These models quantify the variation between and within groups, for example variation between classrooms and within a classroom. A specification for the mixed-effects model is $Y = X\beta + ZU + \epsilon$, where X is a matrix of fixed covariates, Z is a matrix of the random covariates, and $U \sim N(0, \Sigma_U)$ is independent from $\epsilon \sim N(0, \sigma^2_\epsilon)$ (Milliken, Stroup and Wolfinger, 1996). The goal is to estimate the β coefficients, as well as the variance components in Σ_U and σ^2_ϵ. These unknowns are often solved for using maximum likelihood estimaton, and closed form solutions exist only for simple cases. Thus, solution is often estimated using iterative techniques, such as iterative generalized least squares (IGLS).

There are two model-based approaches to incorporating sampling weights in LME models due to Pfeffermann et al. (1998) and Korn and Graubard (2003), PSHGR and KG respectively. Both of these methods focus on *how* to insert weights into the analysis, not *if* or *when* the analysis should include weights. Both utilize a standard practice in survey sampling by starting with the census maximum likelihood equations, writing them in terms of sums, and unbiasedly estimating each of the sums individually using weighted design-based techniques similar to the alternate frequentist model-based method. However, they use different sampling weights and insert them in the analysis in different ways. For simplification, assume that the sampling design first samples clusters and then samples individuals within clusters. To describe the weights used, assume s and t denote subjects, and j represents a cluster, then: w_{js} - the inverse probability that individual s in cluster j is included in the sample, w_j - the inverse probability that cluster j is included in the sample, $w_{s|j}$ - the inverse probability that individual s is included in the sample given cluster j is sampled, and $w_{st|j}$ - the inverse probability that individuals s and t are both included in the sample when cluster j is sampled. The $w_{st|j}$ are referred to as the joint conditional weights.

3.1.1 The PSHGR Approach

PSHGR solve the census likelihood equations using IGLS, and it is within this that they introduce the design-based analysis. To illustrate, suppose they have the random intercept model

$$y_{ji} = x_{ji}\beta + u_j + \epsilon_{ji}, \quad u_j \sim N(0, \tau^2), \ \ \epsilon_{ji} \sim N(0, \sigma^2). \tag{3}$$

From the likelihood equations, a closed form solution for β is available, however not for σ^2 or τ^2. To solve for these, a separate ordinary least squares (OLS) regression is performed where σ^2 and τ^2 correspond to slopes. It is within this secondary OLS and in the closed form solution for β that the sampling weights are inserted. Basically, in solving for these equations, all matrix notation is put in summand format. Then, in sums over clusters indexed by j, the w_j are inserted, and in sums over individuals in clusters, individuals indexed by i and the cluster indexed by j, the $w_{i|j}$ weights are included. This makes the sums design unbiased. This method is referred to as the **weighted unscaled** method. Thus, PSHGR introduce the design-based methodology within the IGLS to provide census instead of sample estimates. Current work is to determine the effect on the likelihood equations of adding the weights in this manner. In addition to providing estimates of the fixed- and random-effects, PSHGR also provide an algorithm for determining the variances of their estimates.

These estimators are consistent for the fixed-effects and the variance components, where consistency requires both the number of clusters and the individuals in a cluster to increase to infinity. PSHGR show that if only the number of clusters increases, then the fixed-effects remain consistent, but not the variance components. These methods will not do well in small area estimation, a branch of survey sampling in which sample sizes within a cluster are held constant.

A downside to the PSHGR method is that the variance components estimators are biased. To adjust for this, they note that multiplying the $w_{i|j}$ weights by scalar does not affect the consistency argument above, but can reduce the bias. Two types of scalars are proposed, and resulting in **weighted scaled 1** and **weighted scaled 2** estimates. The difference between the two scaled estimates is in the way the sample size is estimated. The two weighted scaled estimates are the same if all elements in a cluster have the same weight. PSHGR tentatively recommend the weighted scaled 2 approach for reducing bias caused by informative sampling based on some simulation results, and acknowledge that the simulations represent a limited number of cases.

Criticisms of the PSHGR approaches include 1) they are computationally difficult for someone in the field to use unless implemented into prepackaged software 2) they are not consistent if the number of individuals in a cluster is held constant and 3) the motivation behind their approach is not very clear, and their suggestions lack general guiding principles making it difficult to generalize to other types of models.

3.1.2 The KG Approach

KG also start with the census likelihood equations. They then write the likelihood equations in terms of sums, and insert the weights into the equations to make them design unbiased. For example, the likelihood equations with the weights for a random intercept model are

$$\sum_{j=1}^{K} w_j \frac{1}{\lambda_j} \sum_{i=1}^{M_j} w_{i|j} Y_{ji} - \beta \sum_{j=1}^{K} w_j \frac{M_j}{\lambda_j} = 0 \qquad (4)$$

$$\frac{1}{2}\sum_{j=1}^{K} w_j \frac{1}{\lambda_j^2} \sum_{i=1}^{M_j} w_{i|j} (Y_{ji}-\beta)^2 + \sum_{j=1}^{K} w_j \frac{1}{\lambda_j^2} \sum_{s=1}^{M_j} \sum_{t<s} w_{st|j} (Y_{js}-\beta)(Y_{jt}-\beta) - \frac{1}{2}\sum_{j=1}^{K} w_j \frac{M_j}{\lambda_j} = 0 \qquad (5)$$

where $\lambda_j = \frac{1}{M_j\sigma_a^2+\sigma_e^2}$ and M_j is the size of the population in cluster j. This is referred to as the **joint conditional weighted** analysis, as they use the $w_{st|j}$ weights. They suggest using the jackknife method for survey sampling to compute variances. While developing their technique, they looked for consistency under different asymptotic schemes (Korn, personal communication, December 8, 2003). KG provide simulations with comparable estimators based on the method of moments that perform well when the sample sizes within a cluster are small, though it is not clear that this will be the case for the MLE estimation. KG believe these estimators are not expected to perform well when the number of clusters sampled is small relative to the number of clusters in the population, though the reasoning is not specified. The joint conditional weights, $w_{st|j}$ are more difficult to obtain in practice than the weights needed for the weighted unscaled, weighted scaled 1 and weighted scaled 2 PSHGR analyses.

Criticisms to the KG approaches include 1) they require weights which currently are not calculated routinely, namely $w_{st|j}$ 2) they are difficult to obtain computationally for field workers until they are added to software packages and 3) the authors are not very clear in describing their approach, and it is difficult to see their optimality/evaluation criteria and hence, it can be difficult to generalize.

3.2 Sampling Design/Data Collection Induces Variance Components

The second component of this research is to investigate the use of weights when the sampling design and the model which generated the data do not have parallel variance components, specifically with the use of a model other than an mixed-effects model. Three areas of research apply when the sampling design/data collection induces variance components. The first approach reviews mixed-effects models, with the goal of expanding this to other models. Crossed random-effects are used to account for these variance components in mixed-effects models. Goldstein (1987) introduced these and Raudenbush (1993) and Hill and Goldstein (1998) expanded upon them. Suppose the generating model is $Y_{ji} = \beta_0 + \Sigma_{p=1}^{P} \beta_p x_{pji} + U_{0j} + \Sigma_{q=1}^{Q} U_{qj} z_{qji} + \epsilon_{ji}, \epsilon_{ji} \sim N(0, \sigma_e^2), U_q \sim N(0, \tau_q^2)$. To add a cross effect, say for interviewers, more notation is needed. Let F be the number of interviewers in the data. The random-effect of interviewer f, W_f is added to the model to produce $Y_{ji} = \beta_0 + \Sigma_{p=1}^{P} \beta_p x_{pji} + U_{0j} + \Sigma_{q=1}^{Q} U_{qj} z_{qji} + W_{f(i,j)} + \epsilon_{ji}$ where $f(i,j)$ is the interviewer of individual i in cluster j. There are F new random variables, $W_1, ..., W_F$, which have a normal distribution with mean 0 and variance τ_W^2. Generalizations of this model can include crossed slopes and effects that are both nested and crossed. The full model can be written as $Y = XB + ZU + TW + \epsilon$ where XB represents the fixed effects, ZU represents the random effects, and TW represents the crossed random effects.

The second approach, described by Fienberg and Tanur (1987, 1988), involves embedding an experiment into the survey including models other than mixed-effects models. They explain and build upon the work of Mahalanobis (1946) who embedded an experiment to test for differences between interviewers in a survey on economic conditions of factory workers in India. This work may be expanded to include other sampling design and data collection issues. One aspect of this work is that it necessitates involvement upfront, when the survey is being designed.

A third approach involves estimating equations, as discussed in Binder and Patak (1994). They take the core estimating equation and add additional terms to represent auxiliary data regarding a model. For example, suppose the mean of a population is of interest. Then the core estimating equation is $\int (y - \theta) d\hat{F}_Y(y) = 0$ where $\hat{F}_Y(y) = \Sigma_{i=1}^{n} I(Y_i < y)/n$, n is the number of elements in the sample and $I(Y_i < y)$ is the indicator as to whether $Y_i < y$. However, assume that the mean is equal to $X\beta$ where X is known and β is to be estimated. Then the new estimating equation would be $\int (y - \theta) d\hat{F}_Y(y) + \int (X\beta - \theta) d[F_X(x) - \hat{F}_X(x)] = 0$. This approach can be used towards both finite and superpopulation parameters. This work can be expanded to include, say, a term for data collection and a model for data generation.

3.3 Guidelines For the Use of Weights

The third component to this research is to propose a new set of guidelines and demonstrate them through a series of tutorials. Currently, there are main three groups of guidelines for the use of weights in model based analysis. The first set of guidelines advocated by, for example, Hoem (1989) and Fienberg (1989 suggests that weights should not be used in model based analysis, with the possible exception of when there is informative sampling. The second guidelines, supported by, for example, Kalton (1989), suggests that the weights should always be used, as they help protect against model mis-specification. The third set of guidelines, supported by Lohr and Liu (1994) among others, suggest running two analyses, one with the weights and one without. If the two analyses ”match,” then the use of weights does not matter, otherwise, adjust the model until the weighted and unweighted analyses match. In contrast, DuMouchel and Duncan (1983) propose guidelines based on estimands of interest in fixed-effects regression models, with some using weights, and others not. My research will expand this to different types of model-based analyses.

4 Preliminary Results

4.1 Weights in Mixed-Effects Models

4.1.1 Preliminary Results - Simulations

No existing commercial software computes the PSHGR or KG methods, so I developed a c-program to implement them. The program computes the unweighted analysis, the three PSHGR methods and the KG method, and presents them side by side. I expanded the method proposed in KG to include fixed covariates, as their published method focuses on random intercept models with no covariates. Currently, the program supports only random intercept models with any number of fixed covariates.

To compare the methods with simulations, hierarchical data were generated according to a model ξ to create the finite population, and then the finite population is sampled according to a sampling scheme p. The ξ model is $Y_{ji} = \beta_0 + \beta_1 x_{ji} + U_i + \epsilon_{ji}, U_i \sim \text{Normal}(0, \tau^2), \epsilon_{ji} \sim \text{Normal}(0, \sigma^2)$. The x_{ji} values were created as Normal random variables, with mean one and variance 25. The goal is to estimate β_0, β_1, τ^2 and σ^2 whose true values in the simulation are 1, -2, 0.2, and 0.5. The number of simulated clusters in the population was 300, of which 35 were sampled. The population sizes within each cluster were randomly generated as in PSHGR and ranged from 38 to 147, of which 20 were sampled.

For each simulation, one thousand finite populations were created using the ξ model and each population was sampled independently according to the p-distribution. Three different simulations were conducted, with different p distributions, as in PSHGR. The first simulation contained informative sampling at the cluster level (proportional to the random-effect U_j) and individual levels (individuals with positive ϵ_{ji} being undersampled). The second simulation sampled the groups informatively (as in the first simulation), but the individuals were sampled with simple random sampling. The third simulation was non-informative, with the sampling of groups being proportional to the number of people in the group.

My simulations improved upon the PSHGR/KG simulations in three ways. First, the PSHGR/KG simulations included only random intercept models with no fixed covariates, whereas my simulations included a fixed effect. Second, MSE was computed as a basis of comparison. Third, because of the 1000 iterations in the simulation, a sample distribution for the differences between the different estimands (i.e. subtracting the KG and scaled 2 estimates for β_0) were used for testing significant differences between methods.

The first simulation oversampled clusters with larger random effects and individuals with smaller error terms. The results of this simulation are in Table 2. From this table, we see the unweighted β_0 is borderline biased, corrected in all the weighted analyses. The scaled 2 analysis has the lowest MSE for β_0 (though the differences are very small), and the unweighted has the lowest MSE for β_1. Overall, the weighted approaches appear comparable for the fixed-effects. For τ^2 all estimates are unbiased and for σ^2 the unweighted is biased low. The smallest MSE for τ^2 is the unweighted estimate, and for σ^2 is the scaled 2 estimate. The significant pairwise tests with $\alpha = 0.05$ are in Table 3. Overall, comparing the KG, scaled 1 and scaled 2 estimates, the scaled 2 estimates appear to have the smaller MSE's, but often by a small amount.

The second simulation oversampled clusters with larger random effects. The results of this simulation are in Table 4. In this simulation, the Scaled 1 and Scaled 2 estimators are identical, because the weights of subjects in a cluster are constant. From the table, we see the unweighted β_0 estimate is biased, corrected in all the weighted analyses, and the scaled 1/scaled 2 estimates have the lowest MSE's for β_0, but the difference is very small. The unweighted has the lowest MSE for β_1. Overall, the weighted approaches appear comparable for the fixed-effects. For τ^2 and σ^2, all estimates are unbiased and the unweighted has lowest MSE for both. The significant pairwise tests with $\alpha = 0.05$ are in Table 3. Overall, some of the KG estimates have smaller MSE than the scaled 2 estimates (τ^2), and some of the scaled 2 estimates have smaller MSE than the KG estimates ($\beta_0, \beta_1, \sigma^2$), but the differences are small.

The third simulation was non-informative for clusters and individuals. The results of this simulation are in Table 5. In this simulation, the scaled 1 and scaled 2 estimators are identical, because the weights of subjects in a cluster are constant. From the table, we see that the unweighted estimate always has the smallest MSE. For β_0 and β_1 all the estimates are unbiased, and of the scaled 1, scaled 2 and KG estimates, the KG has the smallest MSE, though the difference is small. For τ^2 and σ^2, the estimates are all unbiased, and of the scaled 1, scaled 2 and KG estimates, the KG has the smallest MSE, though the difference is small. The significant pairwise tests with $\alpha = 0.05$ are in Table 3. While some of these tests showed differences between the scaled 1/scaled 2 estimators and the KG estimators, the differences were very small.

A problem with these simulations was that they did not separate the strengths and weaknesses of the different methods. The simulations mimicked the simulations in PSHGR, but better simulations are necessary to demonstrate the differences in the methods, for example expanding the model beyond random intercepts, varying the percentage of clusters and or individuals within a cluster which are sampled.

For these simulations, the generating model was $Y = \beta_0 + \beta_1 X + U + \epsilon$ which matched the model being estimated. However, for the first and second simulations (those which included informative sampling), the model being estimated did not take into account the informativeness in the sampling design, and thus was not the correct model. In these simulations, the unweighted β_0 estimates (and the σ^2 estimate) were biased, which was corrected for by the weights. Thus, the use of the weights may be appropriate under types of model mis-specification and/or informative sampling. A goal of this research is to better quantify this.

4.1.2 Preliminary Results – Real Data

Next, lessons learned from the simulations are applied to the National Survey of Child and Adolescent Well-Being (NSCAW). NSCAW is a three year longitudinal survey conducted by the Department of Health and Human Services to investigate fundamental questions about the outcomes of abused and neglected children who have come into contact with the child welfare system. NSCAW uses a complex multi-stage sampling strategy (Dowd, Kinsey, Wheeless, Thissen, Richardson, Mierzwa and Biemer, 2001) which first stratifies, then clusters, then stratifies. There is oversampling of subpopulations, such as children entering the child welfare system as infants, families who received child welfare services, and sexual abuse cases. Of interest is the association between the cognitive stimulation of a child and the presence or absence of maternal depression and substance abuse. Based on the sampling design, expert knowledge and exploratory data analysis, the model being estimated includes the variables in Table 6 and is specified as

$$\begin{aligned} CS_{ji} &= \beta_0 + U_j + \beta_{DEPR} * DEPR_{ji} + \beta_{SUBSABU} * SUBSABU_{ji} + \beta_{SEXABU} * SEXABU_{ji} \\ &\quad + \beta_{INFANT} * INFANT_{ji} + \beta_{SERVC} * SERVC_{ji} + \beta_{CGDRACE} * CGDRACE_{ji} \\ &\quad + \beta_{DEPR.SUBSABU} * DEPR_{ji} * SUBSABU_{ji} + \epsilon_{ji}, \end{aligned} \tag{6}$$

where $U_j \sim N(0, \tau^2)$ is a random intercept and $\epsilon_{ji} \sim N(0, \sigma^2)$ is the random error term. The w_{sj} weights are provided with the standard NSCAW data package. Follow-up discussions with the NSCAW designers provided conditional and cluster weights. For simplicity, it was assumed that $w_{st|j} = w_{s|j} w_{t|j}$. The weights in this example add many assumptions to the analysis. The weights are adjusted for a subject's non-response, post-stratification, and times when a specific child welfare agency could not respond for, possibly, months at a time.

The estimates for this model are in Table 7. The standard errors used for the PSGHR approaches are from the derivations in their paper. The standard errors used for the KG method were obtained by a jackknife approximation described Korn and Graubard (1999). The fixed-effects coefficients marked with an asterisk (”*”) were significantly different from zero at the 0.05 level.

From Table 7, we observe that the DEPR, SUBSABU, and SEXABU variables are not significant under any of the methods, the intercept, INFANT, SERVC, and CGDRACE are significant under all the methods, and DEPR*SUBSABU is significant under the unscaled and KG methods, and not the others. The interpretation of this model will be restricted to the variables which were considered significant under any of the five estimation methods, and to the variance components.

Under each simulation, the sampling design did not depend on the fixed-effect variable X and the estimate for β_1 was unbiased. From the NSCAW data, the fixed coefficient estimates (not including β_0) change between 16% (β_{INFANT}) to 50% ($\beta_{CGDRACE}$). From this, it appears that there is informative sampling related to the INFANT, SERVC and CGDRACE variables, though the magnitude of the differences needed for this assertion needs more investigation. The β_0 estimate changes only 5%, so it is not clear if this is related to informative sampling. For the substance abuse/depression interaction, the KG and unscaled estimates two to three times higher than the unweighted scaled 1 and scaled two estimates. It appears that the sampling design is informative with respect to this interaction, but it is difficult to determine if it is significant in the model. We see that the KG estimate for τ^2 is smaller than the unweighted estimate, but the σ^2 estimate for KG is the largest of the weighted estimates, but smaller than the σ^2 unweighted. This is similar to the pattern observed in the informative cluster/informative individual sampling design. In the simulations, observe that the estimates for σ^2 are fairly steady across the second and third simulations where there is no informative sampling of subjects. In the first simulation, with informative sampling of subjects, there are estimates whose confidence intervals for σ^2 do not overlap (unweighted and KG). In the NSCAW data, there are confidence intervals that do not overlap for σ^2.

Pulling this together, it appears that the NSCAW data has informative sampling with respect to both the clusters and the subjects within a cluster. This informative sampling appears related to the INFANT, SERVC and CGDRACE variables. We know from the sampling design, that the INFANT and SERVC variables were used in the oversampling scheme, and it appears that they directly affect the cognitive stimulation scores also. Thus, comparing the different weighted techniques provides us with information regarding possible informative sampling/model miss-specification, however it does not help us determine which weighted analysis is the better one to use. Better simulations are needed to justify these results.

5 Proposed Work

This section outlines work I propose to do to complete the thesis.

1. Investigate two current proposals for incorporating sampling weights to robustly estimate mixed-effects models in the presence of model miss-specification and/or informative sampling, comparing their advantages and disadvantages in various situations.

   First, the c-program which compares the two current proposals (PSHGR and KG) will be expanded beyond random intercept model to allow for random slopes and intercepts. Next, various simulations will be conducted to evaluate the effect of the weights on the robustness of the analysis. The simulations will 1) vary the properties of the sampling design (i.e. informativeness, sampling sizes, etc.), 2) vary the property of the weights (i.e. inverse probability weights or adjusted weights), 3) assume independence on the weights (i.e. $w_{st|j} = w_{s|j}w_{t|j}$), 4) include different types of model miss-specification and 5) vary the model to include different types of random-effects.

2. Address the role of sampling weights when the survey design was not formulated according to a particular statistical model of substantive interest, thus introducing additional variance components, and consider models in addition to mixed-effects models to see how the work generalizes across model-based techniques.

   In expanding beyond mixed-effects models and to another survey with a different complex design, the investigation of weights will be expanded beyond mixed-effects models and the NSCAW survey. Although the first attempt at addressing additional variance components is with mixed-effects models, it will be expanded beyond them. In mixed-effects models, the literature focuses on the generating models which may have crossed random-effects, for example, when students are nested in schools and there is a suspected neighborhood effect. Many neighborhoods feed one school and there are often multiple schools in a neighborhood, so schools and neighborhoods would be crossed. this will be expanded to include crossed effects for sampling design/data collection issues. To incorporate the weights, methods such as PSHGR and KG could be applied.

   Incorporating an experiment into the survey necessitates upfront work when the survey is being designed. For secondary analysts working with large national surveys, this may not be possible. However, if the survey is designed so that it is possible to do, this will be investigated further. The insertion of survey weights is not so clear, and would need further investigation.

   Estimating equations are used in the literature for survey sampling, especially model-based survey sampling. However, often the estimating equations are weighted by the sampling weights to be design unbiased, and then the solution of the weighted estimating equation estimates the parameter of interest. This is considered controversial, as it intentionally uses the weights to compensate for the randomization distribution. Adding different terms to the estimating equations (as described in Section 3.2) is way to use estimating equations without systematic use of the sampling weights.

   With each of the methods above, many questions of interest will be addressed. For example, is there a difference in the insertion of weights if the data collection induces variance components as opposed to the sampling design? What happens when the model (or sampling design or data collection) is miss-specified? Under what conditions do the weights make the analysis more robust? What is the role of weights if there is informative sampling?

3. Apply to data from national surveys

   For the mixed-effects analysis, the NSCAW example presented in Section 4.1.2 will be expanded. From applying what is learned about the weights in the additional simulations, the analysis from the preliminary results will be interpreted to understand if the differences in the weighted analyses are due to informative sampling or model mis-specification. Then model will be updated if needed, and the weights will be used if needed. These changes will aid in the interpretation of the interaction between depression and substance abuse, which is directly relevant to the question of interest.

   To address the role of the sampling weights when the variance components induced by the sampling design/data collection, another survey and model will be used. The National Long Term Care Survey (NLTCS) is administered by the Census Bureau with the goal of examining health and behavioral factors associated with changes in chronic disability and mortality. It is a longitudinal survey started in 1982, with the sixth wave currently in progress. The core aspects of the survey gather information regarding medical conditions of the individual, activities of daily living, support people, nutritional and social activities, cognitive functioning, housing and neighborhood characteristics, health insurance, military service, ethnicity, income, and assets, both of individuals outside and inside of institutions. Sample research areas in the NLTCS include changes in chronic disability in the U.S. Elderly Population, severe cognitive impairment and its impact on the individual and his/her family, effects of disability on medicare and social security expenditures, and age and gender specific effects of nutrition on functioning of U.S. elderly, to name a few. The data from the NLTCS is categorical and a class of models which analyzes cross categorical outcomes, instead of sums of scores, will be investigated.

4. Propose and illustrate guidelines regarding the use of sampling weights in model-based analysis, and develop these recommendations into a series of publishable tutorials that survey to clarify the issues and methodologies for survey analysts.

   In working with two different models, mixed-effects models and a categorical model, the study regarding the weights in model-based analysis will be generalized. Types of guidelines which I would like to make include the conditions under which weights help provide more robustness to the analysis, and consequently, conditions upon which the weights should not be used. This robustness may be against model mis-specification and/or informative sampling.

   A series of tutorials comparing/combining design- and model-based analyses will be developed to help clarify issues and methodologies regarding modeling and weights for survey analysts. These will include the four analysis methods outlined in Section 2 and cover different estimands and sampling schemes, starting with simpler models/designs and moving towards more complex models/designs. These tutorials will be designed to help practitioners understand situations in which the weights should be used, consequences of using the weights when they aren't needed (and vice versa) and examples of how to do their own investigations if there is a specific model of interest to investigate.

Tables and Figures

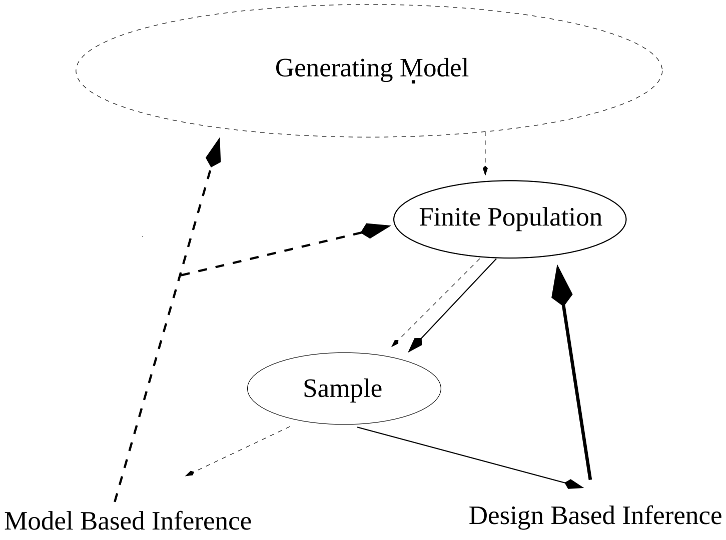

Figure 1: Design-Based (solid lines) and Model-Based (dashed lines) Schematic
Thick Lines Indicate Direction of Inference

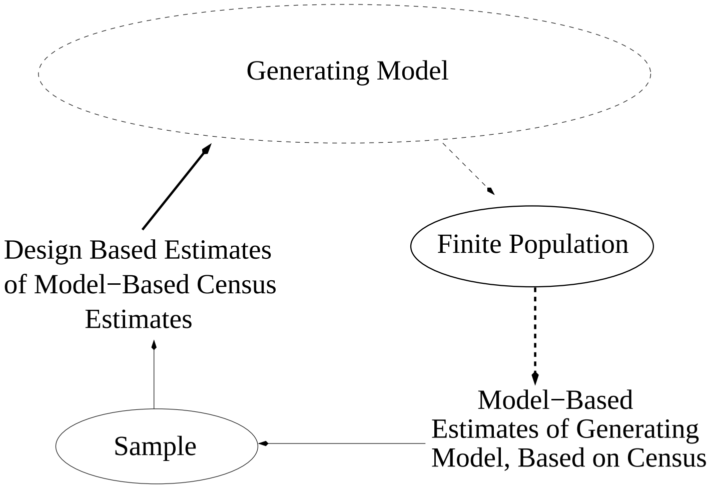

Figure 2: Alternate Frequentist Schematic

	Design-Based	Model-Based
Data Variability	Fixed	Random
Variability definition	Difference across samples	Model error term
Underlying distribution	Randomization distribution	Model distribution
Target of inference	Finite population	Underlying stochastic model or finite population
When sample = census	Parameter known exactly	Model parameter not known exactly Finite population parameter known exactly
Sampling design	Define underlying distribution	Identify dependencies and additional variation

Table 1: Summary of Design- vs. Model-Based Methodologies

Parameter (true value)	Unweighted Estimate (std. dev.) [MSE]	Unscaled Estimate (std. dev.) [MSE]	Scaled 1 Estimate (std. dev.) [MSE]	Scaled 2 Estimate (std. dev.) [MSE]	KG Estimate (std.dev) [MSE]
β_0 (1)	0.87 (0.075) [21.629e-3]	0.99 (0.090) [8.199-3]	0.99 (0.090) [8.211e-3]	0.99 (0.090) [8.183e-3]	0.99 (0.090) [8.198e-3]
β_1 (-2)	-2.00 (5.056e-3) [2.557e-5]	-2.00 (7.433e-3) [5.526e-5]	-2.00 (6.800e-3) [4.619e-5]	-2.00 (6.792e-3) [4.614e-5]	-2.00 (7.430e-3) [5.521e-5]
τ^2 (0.2)	0.17 (0.046) [2.873e-3]	0.20 (0.054) [2.936-3]	0.17 (0.054) [3.683e-3]	0.18 (0.054) [3.245e-3]	0.16 (0.051) [3.937e-3]
σ^2 (0.5)	0.43 (0.024) [5.273e-3]	0.49 (0.035) [1.381e-3]	0.52 (0.034) [1.413e-3]	0.51 (0.033) [1.165e-3]	0.53 (0.037) [2.124e-3]

Table 2: Results from Simulation with Informative Clusters/Informative Subject Sampling

	β_0	β_1	τ^2	σ^2
Simulation 1 informative clusters informative individual	UNWGT < UNSCAL UNWGT < S1 UNWGT < S2 UNWGT < KG		KG <UNWGT KG < UNSCAL KG < S1 KG < S2 S1 < UNSCAL S2 < UNSCAL S1 < S2	UNWGT < UNSCAL UNWGT < S1 UNWGT < S2 UNWGT < KG UNSCAL < KG UNSCAL < S1 S2 < S1
Simulation 2 informative cluster non-informative individual	UNSCAL < UNWGT S1/S2 < UNWGT KG < UNWGT		KG < S1/S2 S1/S2 <UNSCAL KG < UNSCAL	UNSCAL < KG
Simulation 3 non informative cluster non informative individual			KG<S1/S2	UNWGT < KG UNSCAL < UNWGT UNSCAL < KG

Table 3: Significant Differences from Pairwise Simulation Testing

Parameter (true value)	Unweighted Estimate (std. dev.) [MSE]	Unscaled Estimate (std. dev.) [MSE]	Scaled 1 Estimate (std. dev.) [MSE]	Scaled 2 Estimate (std. dev.) [MSE]	KG Estimate (std.dev) [MSE]
β_0 (1)	1.16 (0.076) [32.035e-3]	1.00 (0.086) [7.389e-3]	1.00 (0.086) [7.381e-3]	1.00 (0.086) [7.381e-3]	1.00 (0.086) [7.389e-3]
β_1 (-2)	-2.00 (5.72e-3) [3.256e-5]	-2.00 (6.61e-3) [4.377e-5]	-2.00 (6.19e-3) [3.839e-5]	-2.00 (6.19e-3) [3.839e-5]	-2.00 (6.62e-3) [4.384e-5]
τ^2 (0.2)	0.18 (0.054) [3.008e-3]	0.21 (0.060) [3.709e-3]	0.19 (0.060) [3.707e-3]	0.19 (0.060) [3.707e-3]	0.18 (0.058) [3.626e-3]
σ^2 (0.5)	0.50 (0.028) [0.805e-3]	0.48 (0.030) [1.258-3]	0.50 (0.030) [0.897e-3]	0.50 (0.030) [0.897e-3]	0.51 (0.032) [1.079e-3]

Table 4: Results from Simulation with Informative Clusters/Non-Informative Subject Sampling

Parameter (true value)	Unweighted Estimate (std. dev.) [MSE]	Unscaled Estimate (std. dev.) [MSE]	Scaled 1 Estimate (std. dev.) [MSE]	Scaled 2 Estimate (std. dev.) [MSE]	KG Estimate (std.dev) [MSE]
β_0 (1)	1.00 (0.078) [6.042e-3]	1.00 (0.083) [6.834e-3]	1.00 (0.083) [6.906e-3]	1.00 (0.083) [6.906e-3]	1.00 (0.083) [6.821e-3]
β_1 (-2)	-2.00 (5.50e-3) [2.996e-5]	-2.00 (5.48e-3) [3.012e-5]	-2.00 (5.94e-3) [3.534e-5]	-2.00 (5.94e-3) [3.534e-5]	-2.00 (5.49e-3) [3.023e-5]
τ^2 (0.2)	0.19 (0.055) [3.037e-3]	0.21 (0.057) [3.367e-3]	0.19 (0.058) [3.376e-3]	0.19 (0.058) [3.376e-3]	0.19 (0.055) [3.253e-3]
σ^2 (0.5)	0.50 (0.027) [0.732e-3]	0.48(0.026) [1.057e-3]	0.50 (0.029) [0.856e-3]	0.50 (0.029) [0.856e-3]	0.51 (0.027) [0.781e-3]

Table 5: Results from Simulation with Non-Informative Clusters/Non-Informative Subjects

Variable Name	Variable Meaning
CS_{ji}	The cognitive stimulation score of the child
$DEPR_{ji}$	An indicator of depression in the mother of the child
$SUBSABU_{ji}$	An indicator of substance abuse in the mother of the child
$SEXABU_{ji}$	An indicator as to whether the child was sexually abused
$INFANT_{ji}$	An indicator as to whether the child was under one year
$SERVC_{ji}$	An indicator as to whether the family received child welfare services
$CGDRACE_{ji}$	An indicator if the mother's race was Caucasian
$DEPR_{ji} * SUBSABU_{ji}$	The interaction between $DEPR$ and $SUBSABU$

Table 6: Variables Included in the NSCAW Analysis

Parameter	Unweighted Estimate (std. dev)	Unscaled Estimate (std. dev)	Scaled 1 Estimate (std. dev)	Scaled 2 Estimate (std.dev)	KG Estimate (std.dev)
β_0	6.70* (0.106)	6.93* (0.186)	6.95* (0.232)	6.98* (0.183)	6.89* (0.173)
β_{DEPR}	0.092 (0.099)	0.224 (0.176)	0.138 (0.185)	0.241 (0.197)	0.220 (0.172)
$\beta_{SUBSABU}$	-0.099 (0.254)	0.446 (0.293)	0.419 (0.369)	0.364 (0.275)	0.460 (0.319)
β_{SEXABU}	0.323 (0.252)	-0.017 (0.282)	-0.443 (0.378)	-0.444 (0.338)	-0.016 (0.234)
β_{INFANT}	-0.586* (0.092)	-0.504* (0.109)	-0.553* (0.144)	-0.545* (0.116)	-0.521* (0.105)
β_{SERVC}	-0.363* (0.104)	-0.467* (0.198)	-0.373* (0.159)	-0.472* (0.139)	-0.432* (0.207)
$\beta_{CGDRACE}$	0.607* (0.081)	0.487* (0.180)	0.449* (0.177)	0.417* (0.144)	0.494* (0.186)
$\beta_{DEPR.SUBSABU}$	-0.656 (0.364)	-1.653* (0.573)	-0.577 (0.456)	-0.680 (0.391)	-1.638* (0.656)
τ^2	0.085 (0.035)	0.339 (0.086)	0.111 (0.086)	0.170 (0.064)	0.064 (0.095)
σ^2	2.066 (0.096)	1.268 (0.107)	1.408 (0.105)	1.341 (0.096)	1.563 (0.134)

Table 7: Results on NSCAW data

References

Dowd, K., Kinsey, S., Wheeless, S., Thissen, R., Richardson, J., Mierzwa, F., and Biemer, P. (2001). *National Survey of Child and Adolescent Well-Being (NSCAW) Wave 1 Data File User's Manual, Restricted Release*. Ithaca, New York: National Data Archive on Child Abuse and Neglect, Cornell University.

DuMouchel and Duncan (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. *Journal of the American Statistical Association*. 78, 535-543.

Fienberg, S. (1980). The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey. *Statistician*. 29, 313-350.

Fienberg, S. (1989). Modeling Considerations: Discussion from a Modeling Perspective. In *Panel Surveys*, Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. P. (eds), 512-539. New York, New York: Wiley.

Fisher, R. A. (1935). *Design of Experiments*. Edinburg. Oliver & Boyd.

Gelman, Carlin, Stern and Rubin (1995). Bayesian Data Analysis. Chapman & Hall. New York.

Godambe, V. P. (1955). A Unified theory of Sampling form Finite Populations. *Journal of the Royal Statistical Society B*. 17, 269-278.

Godambe, V. P. (1966). A New Approach to Sampling from Finite Populations. *Journal of the Royal Statistical Society B*. 28, 310-328.

Goldstein, H. (1987). Multilevel Covariance Component Models. *Biometrika*. 74, 430-431.

Graubard, B. and Korn, E. (1996). Modeling the Sampling Design in the Analysis of Health Surveys. *Statistical Methods in Medical Research*. 5, 263-281.

Hansen, M., Madow, W., and Tepping, B. (1983). An Evaluation of Model-Dependent and Probability Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association*. 78, 776-793.

Hill, P. and Goldstein, H. (1998). Multilevel Modeling of Educational Data with Cross-Classification and Missing Identification for Units. *Journal of Educational and Behavioral Statistics*. 23, 117-128.

Hoem, Jan M. (1989). The Issue of Weights in Panel Surveys of Individual Behavior. In *Panel Surveys*, Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. P. (eds), 512-539. New York, New York: Wiley.

Holt, D., and Smith,T.M.F. (1979). Post-Stratification. *Journal of the Royal Statistical Society A*. 142, 3346.

Horvitz and Thompson (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*. 47, 663-685.

Kalton, G. (1968). Invited Discussion to Smith, T. M. F., Present Position and Potential Developments: Some Personal Views: Sample Surveys. *Journal of the Royal Statistical Society A*. 147, 208-221.

Kalton, G. (1989). Modeling Considerations: Discussion from a Survey Sampling Perspective. In *Panel Surveys*, Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M. P. (eds), 575-587. New York, New York: Wiley.

Korn E. and Graubard, B. (1999). Analysis of Health Surveys. Wiley. New York.

Korn, E. and Graubard, B. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society B*. 65, 175-190.

Little, R. (1991). Inference with Survey Weights. *Journal of Official Statistics*. 7, 405-424.

Little, R. (1993). Post-Stratification: A Modeler's Perspective. *Journal of the American Statistical Association*. 88, 1001-1012.

Little, R. (2004). To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association*. 99, 546-556.

Lohr, S. (1999). *Sampling: Design and Analysis*. New York, New York: Duxbury Press.

Lohr, S. and Liu, J. (1994). A Comparison of Weighted and Unweighted Analyses in the National Crime Victimization Survey. *Journal of Quantitative Criminology* 10, 343-360.

Milliken, G., Stroup, W. and Wolfinger, R (1996). *SAS System for Mixed Models*. SAS Publishing.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*. 97, 558-625.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International Statistic Review*. 61, 317-337.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *Journal of the Royal Statistical Society B*. 60, 23-40.

Raudenbush, Stephen (1993). A Crossed Random Effects Model for Unbalanced Data with Applications in Cross-Sectional and Longitudinal Research. *Journal of Educational Statistics*. 18, 321-349.

Royall, R. (1968). An Old Approach to Finite Population Sampling Theory. *Journal of the American Statistical Association*. 63, 1269-1279.

Royall, R. (1976). Likelihood Functions in Finite Population Sampling Theory. *Biometrika*. 63, 605-614.,

Sarndal (1978). Design-Based and Model-Based Inference in Survey Sampling. *Scandinavian Journal of Statistics*. 5, 27-52.

Sarndal, Swensson and Wretman (1992). Model Assisted Survey Sampling. Springer, New York.