Approaches for Analyzing Survey Data: a Discussion

David Binder[1], Georgia Roberts[1]
Statistics Canada[1]

Abstract

In recent years, an increasing number of researchers have been able to access survey microdata files. These researchers perform various analyses to summarize the data and to describe relationships in a target population. Many of these researchers use analytic software without having a good understanding of the statistical underpinnings behind the methods being applied. Some of the issues facing the researchers include understanding the survey weights, understanding informative sampling, using variances that are model-dependent, incorporating survey design information into the modeling process, and integrating data from more than one survey. In this paper we discuss a framework within which these issues may be discussed.

Keywords: Complex survey data, Survey weights, Variance estimation, Survey integration, Model-design-based randomization.

1. Introduction

Data analysis is the process of transforming raw data into usable information. This process involves many important steps, including identifying an issue, asking meaningful questions, developing answers to these questions through examination and interpretation of data and, finally, communicating the results. In recent years, many more researchers have gained access to rich sources of survey microdata and have been asking about appropriate methods for examining and interpreting such data. They know that survey data are complex due to the stratification, clustering and unequal selection probabilities used to select the sample and also due to nonsampling problems such as coverage and nonresponse. They want to know whether and how such complexities should be accounted for when they are interested in investigating a variety of questions about a population - where, sometimes, that population is finite, and, other times, it is infinite. For a general discussion of this topic, see Korn and Graubard (1995) and Graubard and Korn (2002). The purpose of this paper is to propose a framework within which many of these researchers' questions may be discussed. For the remainder of this paper, we will restrict the word "analysis" to refer to the steps in the data analysis process that are involved with the examination and interpretation of the data.

In choosing an appropriate analysis method for survey data, the first question that needs to be addressed is what the target population for the analysis is. In Section 2 of this paper we will define and discuss both finite and infinite target populations and will illustrate their difference through some examples. We will then, in Section 3, discuss the principles for making statistical inferences for the two types of target populations. We will follow this, in Section 4, by an explanation of the most common approaches to analysis of survey data and provide some arguments for choosing a design-based approach when a researcher wishes to estimate and make inferences about model parameters. Finally, in Section 5, we will illustrate the principles and approaches that we are proposing through the examination of questions related to the integration of data from more than one survey in a single analysis. Some concluding remarks are given in Section 6.

2. Target Population of an Analysis

When a researcher begins his analysis, his first step is to specify his target population. The target population is the population about which the researcher wishes to make conclusions. It could vary with the issue being studied, even if the same survey is being used. It also usually differs from - and may not even overlap with - the survey population, which consists of the finite set of all units that are eligible for selection through the frame and survey design being used.

In this paper, we find it useful to categorize target populations by whether they are finite or infinite. Some properties of each category are described in the following two subsections.

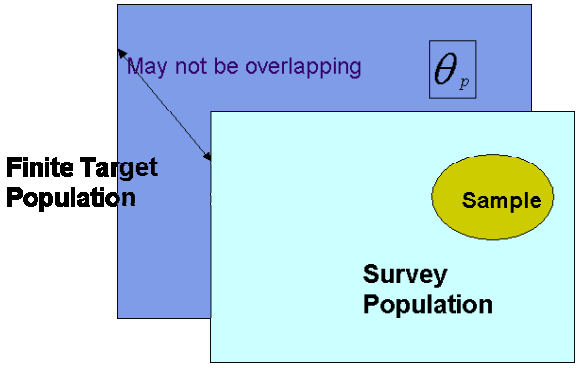
2.1 Finite Target Population

When his target population is finite, the quantities of interest to the researcher are generally characteristics of that finite population. These are characteristics such as a population average or population prevalence.

When planning and implementing a survey and preparing the resultant data files, the survey producer also has a target population in mind, which may or may not

coincide with the researcher's target population. While the survey producer's target population is finite, like the survey population, these two populations usually differ, as seen in Figure 1. In the case of a household telephone survey, for example, the survey population would lack any individuals in households without a telephone, even though these people could be in the survey producer's target population. The survey producer usually provides weights in his data files to allow estimation of characteristics of his finite target population. These weights contain adjustments for known differences between the survey producer's survey and target populations. If the researcher's target population differs from the survey producer's target population, adjustments to the weights provided by the survey producer may be required to account for these differences.

Figure 1. Finite Target Population and Survey Population



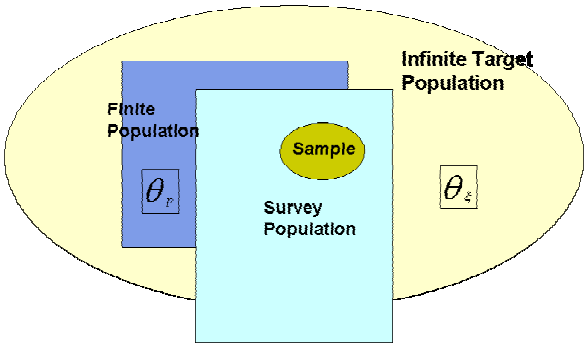
An example of a research question related to a characteristic of a finite target population is the following: "Was there a difference in 2002 between Ontario and Quebec organic farmers in average expenses per acre to grow tomatoes?" To study such a question, the researcher might have access to the data from a 2002 cross-sectional survey of Canadian farmers where questions were asked about organic farming techniques used that year for various crops. The researcher's target population is a domain in the finite population targeted by the survey provider.

2.2 Infinite Target Population

A researcher's target population is generally said to be infinite when the values of variables for this population are thought to have been generated by a statistical model. The quantities of interest to the researcher are characteristics of the model, such as the model parameters. Consider, for example, the problem of investigating whether obesity is a risk factor for arthritis, controlling for age and sex. In this case the researcher may have a logistic model in mind and be particularly interested in the coefficient of the obesity variable. The researcher is not confining his target population to any finite group at a fixed point in time, but may feel that the logistic model approximately describes the relationships among the variables involved during the past 15 years in western cultures, for example. Thus, his target population could be considered to be infinite. Suppose the researcher had used a 1995 American health survey as his data source for fitting and testing his model. It would seem reasonable to presume that the researcher's logistic model could have generated the values of the variables involved for a finite population such as the finite population targeted by the providers of the data for that health survey.

While the quantities of interest to the researcher are parameters of a model generating an infinite population, there are finite population parameters associated with these quantities of interest. In the case of the logistic model described above, the finite population parameters associated with the model coefficients could be the estimates of these coefficients when all the values from the full finite population are available. Such estimates are descriptive parameters of the finite population and frequently are useful summary statistics in their own right. In Figure 2 we illustrate the relationships among the various quantities when the target population is infinite. In this figure, θ_ξ represents the quantities of interest in the infinite target population, whereas θ_p represents the associated finite population quantities.

Figure 2. Infinite Target Population



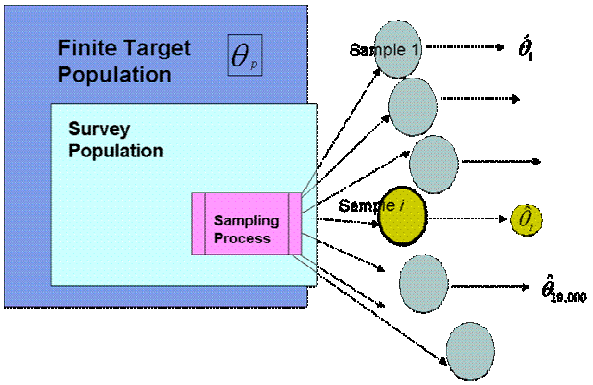
3. Principles for Making Statistical Inference

For statistical inferences, a researcher is interested both in what he observed and in what he did not observe. Of primary interest is the distribution of estimates under hypothetical random repetitions. The distribution of these estimates depends on whether or not a statisti-

cal model is presumed to have generated the values of a finite population, and the properties of the model. As well, the distribution of the estimates may or may not be affected by the sample design.

Consider, first of all, the case of a finite target population where no statistical model is presumed to have generated the finite population and where the only randomization is the design-based randomization. This case is illustrated in Figure 3. Here, the characteristic of interest is a descriptive parameter of the finite population represented by θ_p. Through the sampling design for the survey, sample i is selected and the estimate of θ_p derived from this sample is denoted by $\hat{\theta}_i$. However, it is possible that, under the sampling design used, a large number of samples different from sample i could have been chosen, each of them leading to their specific estimate of θ_p. The distribution of these different possible estimates is what may be called the design-based sampling distribution of the estimate. This is the basis for design-based inferences.

Figure 3. Design-based Randomization



Let us now turn to the case of an infinite target population where the values of variables for this population are described through a model and it is a characteristic of the model, say θ_ξ, that is of primary interest to the researcher. Model-based inferences are based on the sampling distribution of the estimates of that characteristic due to different samples being drawn directly from that model. This is illustrated in Figure 4.

The final case that we wish to present is still the case of the infinite target population where the values of variables for this population are thought to have been generated by a statistical model and it is the characteristics of the model that are of primary interest to the researcher. However, we want to explicitly account for the presumption that the model could have generated the values of the variables in the finite population from which the survey sample was drawn. In this situation, our focus is on the distribution of the estimates of the model parameters of interest, and we want to take account of the variability implied by the model as well as the variability implied by the survey design. This case is called model-design-based randomization and is illustrated in Figure 5. We feel that this is the randomization framework under which many questions related to appropriate analysis methods for survey data could be best explored. For a more rigorous treatment of the asymptotic theory in the design-model-based framework, see Rubin-Bleuer and Schiopu-Kratina (2005).

Figure 4. Model-based Randomization

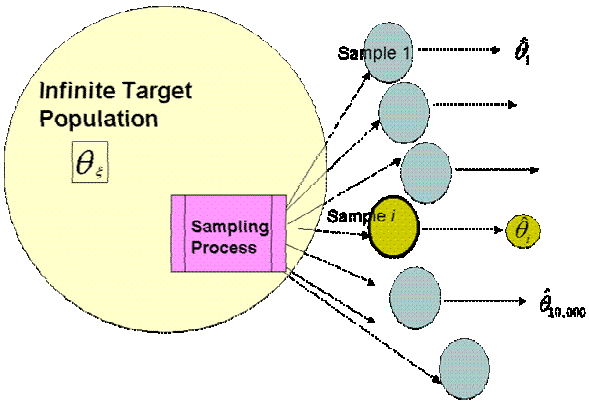
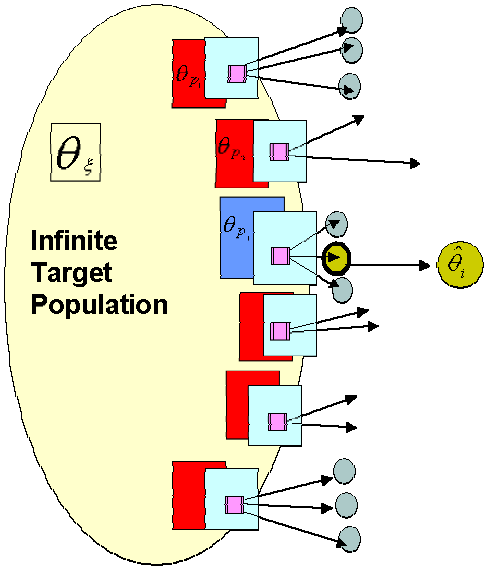


Figure 5. Model-design-based Randomization



In summary, if we let θ represent the characteristic of interest (which could be θ_ξ or θ_p) and if we let $\hat{\theta}$ be

its estimator, then the distribution of $\hat{\theta}$ is the distribution of the different conceptual values of this estimator, depending on the randomization assumptions that have been made: design-based, model-based or model-design-based. This implies, for example, that the expected value of the estimator is

$$E_\infty = \lim_{k\to\infty} \sum_{i=1}^{k} \hat{\theta}_i / k \,,$$

where $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ are k independent draws from the distribution. The bias of $\hat{\theta}$ is then the difference between this expected value and the target parameter. Also, the variance of $\hat{\theta}$ is

$$V_\infty = \lim_{k\to\infty} \sum_{i=1}^{k} (\hat{\theta}_i - E_\infty)^2 / k \,.$$

Both the target population and the randomization assumptions matter when it comes to the values taken by these quantities.

3.1 Informativeness and Ignorability

When variability due both to the model and to the survey design is being considered, two concepts encountered in the literature are informativeness and ignorability. See Pfeffermann (1993) for some discussion of these.

The generation of the observed sample is actually a two-phase process, where at the first phase the finite population is generated according to the model and at the second phase the sample is drawn according to the survey design. When the sample can be assumed to have been generated directly from the model (without this affecting the distribution of the sample variable values), the sampling is said to be not informative. Otherwise it is informative. Simple random sampling designs are noninformative. For more complex sampling plans, whether or not the sampling is informative will depend on the validity of the model assumptions for the observed sample. The concept of informativeness is illustrated in Figure 6.

Next, consider a particular analysis of the data generated from this two-phase process. If a model-based method of inference for the analysis is valid under the two-phase model-design-based randomization process, the sampling is said to be ignorable for that analysis. Otherwise it is nonignorable. For example, when fitting a linear model using ordinary least squares regression estimation, if the actual model residuals are correlated within sampled clusters in a cluster sample, the sample design is nonignorable if the intra-cluster correlation is not properly taken into account. The concept of ignorability is illustrated in Figure 7 for inferences about the model parameter, θ_ξ. It follows that noninformative sampling is ignorable for all analyses (Binder and Roberts, 2001). Some research has been done on diagnostics for ignorability (see, for example, Fuller (1984)).

Figure 6. Non-informative Sample Design

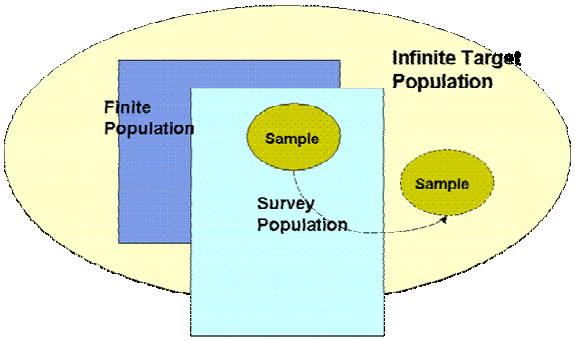
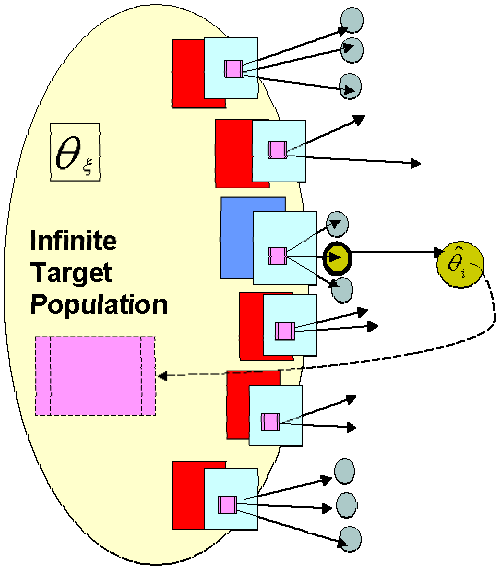


Figure 7. Ignorable Sampling



4. The Most Common Approaches to Analysis

The two approaches commonly used for analyzing survey data are the following:

(a) Design-based: This is the most commonly used approach for estimating finite population quantities for large-scale surveys, and is, as discussed below, also often appropriate when making inferences about model parameters. In this approach, the only source of randomness explicitly accounted for is that due to the survey design. Survey weighting is used to produce esti-

mates of unknown finite population quantities – which are the descriptive quantities of interest in the case of a finite target population and are related to the model quantities of interest in the case of an infinite target population. Design-based variance measures the variability among estimates from possible samples selected by the same design from the same finite population. There are a variety of methods for obtaining design-based variance estimates.

(b) Model-based: This approach, which is generally used when the quantities of interest are the parameters of a model, assumes that all randomness is expressed explicitly in the model. It is thus possible that a model for the infinite population will need modification so that it details the impact of the survey design on the variables being described in the sample taken. Classical non-survey approaches are used to fit the model, estimate variances and make inferences.

4.1 Why Take a Design-based Approach

When the target population is infinite and the quantities of interest are parameters of a model generating values of the variables in a finite population, we contend that model-design-based randomization can serve to explain how the survey data were generated. However, we feel that, for a great number of problems studied by researchers, a pure design-based approach can still lead to valid inferences in the model-design-based randomization framework. There are several reasons for this.

First of all, under model-design randomization, a design-based approach gives valid inferences for model parameters when the mean model is approximately correct for the infinite population and when sampling fractions are small. Obviously,

$$\hat{\theta}_p - \theta_\xi = (\hat{\theta}_p - \theta_p) + (\theta_p - \theta_\xi) \ .$$

Thus, if $E_p(\hat{\theta}_p) \approx \theta_p$ and $E_\xi(\theta_p) \approx \theta_\xi$, then

$$E_{\xi p}(\hat{\theta}_p - \theta_\xi) \approx 0 \ .$$

Also,

$$V_{\xi p}(\hat{\theta}_p - \theta_\xi) \approx V_\xi(\theta_p) + E_\xi V_p(\hat{\theta}_p)$$
$$= O(1/N) + O(1/n).$$

If the sampling fraction, *n/N*, is small,

$$V_{\xi p}(\hat{\theta}_p - \theta_\xi) \approx E_\xi V_p(\hat{\theta}_p) \ ,$$

and using $\hat{V}_p(\hat{\theta}_p)$ will give valid model-design-based inferences about θ_ξ .

Secondly, researchers – particularly secondary users of the data – may not know enough about the design to completely model its impact. Even if a researcher does know the design well, suitable design variables may not exist on the data files provided for analysis for inclusion in a parsimonious model. Thus, appropriate modification of a model to explain the survey data may not be feasible and thus a design-based approach may make more sense.

Finally, a researcher may not want design variables in his model since inclusion of these variables could change the interpretation of other model parameters (see, for example, Chambers (1986)). Using the form of the model that generates the infinite population, plus design-based methods to implicitly account for the impact of the survey design on the model holding in the sample thus may seem like a more palatable option.

It should be noted that a pure design-based approach would not be valid under model-design-based randomization when sampling fractions are not small. However, in this case, the model-design-based framework could point to appropriate corrections to the design-based variance estimates.

5. Applying These Principles and Approaches to Integrating Data From More Than One Survey

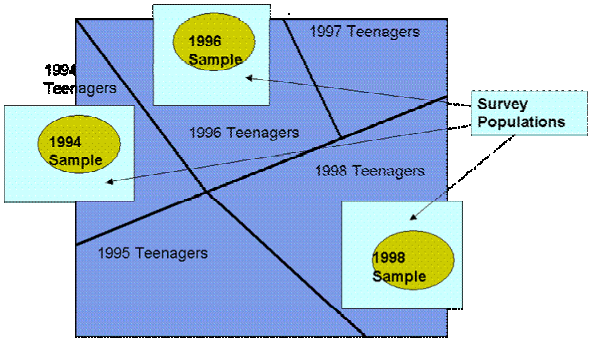
As data are being collected and are being made accessible to researchers from an increasing number of surveys, the researchers are noting that comparable variables of interest are available from more than one survey source. It is often the case that the sample sizes for the problem that they wish to study are small in each of the survey sources. Of interest to these researchers is whether and how to perform the analysis by integrating the data from more than one survey.

5.1 Integrating When Target Population is Finite

Let us start with the situation where the quantity of interest is a descriptive parameter that is a characteristic of a finite population. The quantity of interest could be, for example, the prevalence rate of a disease or the proportion of smokers in a population.

In Figure 8, we illustrate a complex case where teenagers were sampled in 1994, 1996, and 1998. However, the target population of interest to the researcher includes all teenagers in the years 1994 to 1998, so that teenagers in 1995 and 1997 are also part of the researcher's target population. Note that the population of all teenagers in the years 1994-1998 is a conceptual one, since it never exists at any single point in time. Note also that persons who were teenagers in more than one year are considered here as different units in the conceptual finite population.

Figure 8. Integrating with Finite Target Populations



In the case described here, and in many other situations, the question that arises is whether it makes sense to integrate the data from more than one survey. Such integration could be considered when either of the following two conditions apply:
(i) if the researcher's target population is the combination of the finite populations targeted by the survey producer for the different surveys (i.e., each finite population is like a super-stratum). In this case, the quantity of interest need not be assumed to be constant over the different super-strata, although whether or not this is true could influence the choice of approach to integration;
(ii) if the researcher's target population is a bigger population than the combined finite populations targeted by the survey producers, as in our example above. In this case, some assumptions about the relationship between the quantities of interest in the populations that were not sampled with the quantities of interest in the populations that were sampled would need to be made. For example, one might assume that for the population illustrated in Figure 8 the average smoking rate for teenagers in the years 1994-1998 is similar to the average over only the years 1994, 1996, and 1998. Alternatively, for some other characteristic, such as prevalence rate for some health condition, one might assume that the characteristic of interest is constant, or has a constant linear trend, over all the years in the researcher's target population.

In the next two subsections, we describe the two broad choices for integrating the data.

5.1.1 Separate Approach to Integration

The first broad choice for integrating the data would be to estimate the parameter from each data source separately and then to combine the estimates through averaging. Before proceeding, the researcher should perform some preliminary work.

First of all, he should check on the assumption of equality of the parameter across the different finite populations. This confirmatory work could involve some formal statistical testing and also background investigation into the subject matter. (The power of the statistical tests may not be high if the sample sizes from each survey are low.)

Secondly, he should consider the meaning of the average of estimates if the parameters are unequal, and determine whether, in such a case, the average would have relevance to his research.

As well, he should consider whether a weighted average, rather than a simple average, would have more advantages for his particular research. The large body of research into the topics of population-size-adjusted or design-effect-adjusted weighting could help with this decision. However, it is important to note that "optimal" methods for weight adjustments may depend on knowing the variances or design-effects of an estimate, and these variances are often estimated from data based on small sample sizes.

When the surveys are independent, it is usually feasible to construct estimates of the variances for the estimator using a separate approach. On the other hand, when the surveys are not independent, the correlation between surveys will need to be accounted for in the variance estimates.

5.1.2 Pooling Approach to Integration

As a second approach to integration, the researcher could pool the data from the different surveys, considering the data from each as being from a different superstratum, and then treat the data as if from a single survey. However, before proceeding, there are again some things to consider.

The researcher should do some confirmatory work regarding an assumption of equality of the parameter across the superstrata. He should consider the meaning of the pooled estimate if equality is not true. (For example, does he actually want an estimate of the prevalence rate in the pooled populations if the prevalence rates within the different populations are not the same?)

He could also consider whether doing weight rescaling within each data source would be advantageous. For example, he could explore whether it lead to a more efficient estimate. However, in the situation of unequal parameters in the different finite populations, he

would need to consider whether the rescaled estimate would make sense.

As in the case of a separate approach, it is usually feasible to construct estimates of variances when a pooled approach is used.

It should be noted that only under specific conditions would the two approaches – pooled and combined – give the same point estimate (even when estimating the same quantity).

5.2 Integrating When Target Population is Infinite

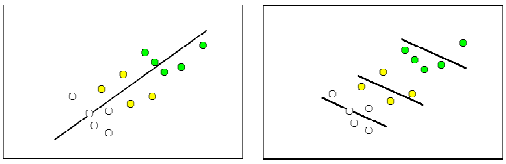
We now turn to the situation where the quantities of interest are parameters of a model describing an infinite population.

It would seem feasible for a researcher to consider integrating the data from more than one survey if the statistical model (which describes an infinite population) could be presumed to have generated the values of each of the finite populations targeted by the survey producers for the different surveys under consideration for integration. Furthermore, the model could – and probably should – contain parameters particular to each finite population.

As is the case for a descriptive parameter of a finite population, either pooling or combining are possible approaches for integrating the data from the different surveys.

However, for the infinite population, where modeling is involved, the pooling approach has some distinct advantages. When pooling, it is generally straightforward to allow for and to test for inequalities in parameters among the different finite populations presumed to have been generated by the model. Consider, for example, the simple situation displayed in Figure 9, where three different surveys collected information on the same two variables and where the model of interest to the researcher posited a linear relationship between the two variables. If the researcher pooled the data from the three surveys and fitted a linear model without consideration of the source of each data point, his estimated line would have had a strong positive slope, as shown on the left of Figure 9. If, however, he allowed for different slopes and intercepts for the different data sources in his model for the pooled data, his estimated lines would have the form shown on the right of Figure 8. It appears as if the lines are parallel, but with a negative slope. Further investigation by the researcher reveals that the negative linear relationship between the two variables made sense and that the difference in the locations of the lines for the three finite populations presumed to have been generated by the model could be attributed to a survey effect, such as mode effect, of which the researcher had not been previously aware.

Figure 9. Fitting Linear Models Using Integrated Surveys



6. Conclusions

There is controversy about using a design-based approach for estimating model parameters. We feel that the issues raised in this controversy can be discussed and clarified in a model-design-based framework. As well, as shown in this paper, use of this framework will identify the situations where a pure design-based approach makes sense. In these discussions, the notion of the appropriate target population is important.

References

Binder, David A. and Roberts, Georgia R. (2001), "Can Informative Designs be Ignorable?" *Newsletter of the Survey Research Methods Section, Issue 12,* American Statistical Association.

Binder, David A. and Roberts, Georgia R. (2003), "Design-based and Model-based Methods for Estimating Model Parameters," in *Analysis of Survey Data,* (eds. R.L. Chambers and Chris Skinner) Wiley, Chichester, pp. 29-48.

Chambers, R.L. (1986), "Design-Adjusted Parameter Estimation," *Journal of the Royal Statistical Society, Series A, 149,* pp. 161-173.

Fuller, Wayne A. (1984), Least Squares and Related Analyses for Complex Survey Designs. *Survey Methodology, 10,* pp. 97-118.

Graubard, Barry I. and Korn, Edward L. (2002), "Inference for Superpopulation Parameters Using Sample Surveys," *Statistical Science, 17,* pp. 73-96.

Korn, Edward L. and Graubard, Barry I. (1995), "Analysis of Large Health Surveys: Accounting for the Sampling Design," *Journal of the Royal Statistical Society, Series A, 158,* pp. 263-295.

Pfeffermann, Danny (1993), "The Role of Sampling Weights When Modeling Survey Data," *International Statistical Review, 61,* pp. 317-337.

Rubin-Bleuer, Susana, and Schiopu-Kratina, Ioana, (2005), "On the Two-Phase Framework for Joint Model and Design-Based Inference," *Annals of Statistics, 33,* pp. 2789–2810.