

# HOW IMPORTANT IS THE INFORMATIVENESS OF THE SAMPLE DESIGN?

David A. Binder, Milorad S. Kovacevic and Georgia Roberts<sup>1</sup>

## ABSTRACT

Typical complex sample designs lead to informative samples, which means that the distribution of a variable in the sample is different from its distribution in the population. To determine and measure the impact of informativeness, we compare design-based and model-based variances of estimated parameters (as well as the estimated parameters themselves) in a logistic model under the assumption that the postulated model is true. For an appropriate modelling of data from informative samples we consider two possibilities: to use design-based inference about the model parameters or to use model-based inference. We propose a new bivariate approach for assessing the informativeness of a sample design on data; this approach accounts for effects on point estimates and on their standard errors. A large simulation study, based on generating a population under the postulated model, using parameter estimates derived from the National Population Health Survey (NPHS), allows us to detect and to measure the informativeness, and to compare the robustness of studied approaches.

KEY WORDS: Design-based, model-based, mixed-model approach, informative clustering, power of test

## RÉSUMÉ

Les plans typiques de sondage complexe conduisent à des échantillons informatifs, c'est-à-dire que la distribution d'une variable dans l'échantillon est différente de sa distribution dans la population. Afin de déterminer et de mesurer l'impact de l'informativité, on comparera les variances des paramètres estimés (de même que les paramètres estimés) fondées sur le plan à celles fondées sur le modèle, dans un modèle logistique sous l'hypothèse que le modèle formulé est vrai. Pour évaluer l'adéquation de la modélisation des données provenant d'échantillons informatifs, nous considérerons deux façons de faire: utiliser une inférence fondée sur le plan pour les paramètres du modèle ou utiliser une inférence fondée sur le modèle. Nous proposons une nouvelle approche bivariable pour évaluer l'impact de l'informativité du plan de sondage qui tient compte des effets sur les estimations et leur écart type. Une étude par simulation d'envergure, basée sur la génération d'une population sous un modèle postulé, utilisant des paramètres estimés dérivés de l'ENSP, nous permet de détecter la présence d'informativité, de la mesurer, et de comparer la robustesse des deux approches retenues.

MOTS CLÉS: selon le plan, selon le modèle, selon le plan-modèle, mise en grappes informative, puissance de test

## 1. INTRODUCTION

Informativeness of a sample is a model concept. If the distribution of the sampled units is different from the distribution that would be obtained by sampling directly from the model, then the sampling is said to be **informative**. The design is said to be **ignorable** for a particular analysis if it has the property that the results of the analysis are not affected by the informativeness of the sample design. All non-informative designs lead to ignorability, but not vice versa (Binder and Roberts, 2001).

Some analysts fit the same model to survey data using both a design-based and a model-based approach, and if the point estimates of the model coefficients are similar under the two approaches, they make the conclusion that the sampling was not informative, and carry on with a model-based approach. However, the design-based and model-based point estimates can be similar even when the assumptions about the model distribution are incorrect for the sample. Thus, when the point estimates are similar, but the **estimates of the design-based variances** are not close to the **estimates of the model-based variances**, this could be an indication that the sampling is “informative”, particularly when sample size is large. If the preference is still to take a model-based approach, then the model should be modified to ensure that the sampling distribution for the sampled units is valid under the model. (This may be difficult to achieve, however.)

---

<sup>1</sup> David A. Binder, Statistics Canada, 17J R.H.Coats Building, Tunney's Pasture, Ottawa, ON, K1A0T6, [dbinder49@hotmail.com](mailto:dbinder49@hotmail.com), Milorad S. Kovacevic, Statistics Canada, 17J R.H.Coats Building, Tunney's Pasture, Ottawa, ON, ON, K1A0T6, [kovamil@statcan.ca](mailto:kovamil@statcan.ca), Georgia Roberts, Statistics Canada, 17J R.H.Coats Building, Tunney's Pasture, Ottawa, ON, ON, K1A0T6, [robertg@statcan.ca](mailto:robertg@statcan.ca)

Our goal in this paper is to investigate ways for assessing whether the survey design has an impact on the substantive conclusions from an analysis. In this regard, we compare the standard design-based method of analysis to some alternative frequently-used methods: standard model-based, model-based using standardized weights, and mixed models with random effects for clustering. We proceed by means of a simulation study in which we generate a finite population where our posited model is completely satisfied. Through stratification and clustering of the outcomes, we control for the informativeness in the samples drawn from the finite population. We then assess and compare the impact of the informativeness of the sampling design over the different methods.

Section 2 provides a description of the simulation study. The assessment measures used and the results obtained for assessing the impact of informativeness on point and variance estimates are presented in Section 3. Section 4 contains details of an investigation of the impact of informativeness on power and size of the tests. We propose an approach for assessing the informativeness in sample data in Section 5. Some concluding remarks are given in Section 6.

## 2. SIMULATION STUDY

### 2.1 Superpopulation (Model)

In order to have an empirical assessment of the impact of the informativeness of a sampling design on the analysis of survey data we carried out an extensive simulation study. We simulated a model of the relationship between *the loss of independence among seniors (LOSS)* and several factors associated with their health status and habits, motivated by a model fitted to data from the first two cycles (1994/95 and 1996/97) of the Canadian National Population Health Survey (NPHS), as presented in Martel, Bélanger and Berthelot (2002).

The model that we simulated expresses the probability of an independent senior losing his/her independence as a function of the senior's sex, age, body mass index, presence of chronic diseases and smoking habits. The model has the form:

$$\begin{aligned} \text{logit}(\mathbf{LOSS}) = & \beta_0 + \beta_1 * \mathbf{SEX} + \beta_2 * \mathbf{AGEGR} + \beta_3 * \mathbf{UNDERWGT} \\ & + \beta_4 * \mathbf{OVERWGT} + \beta_5 * \mathbf{CHRDS} + \beta_6 * \mathbf{SMOK} \end{aligned} \quad (2.1)$$

All variables in the model are binary: *LOSS* (1, if person loses his independence within the two-year period studied, 0 otherwise), *SEX* (0 for women, 1 for men), *AGEGR* (0 for age in [65,75), 1 for age 75+), *UNDERWGT* (1 for  $BMI \leq 18.5$ , 0 otherwise), *OVERWGT* (1 for  $BMI \geq 25$ , 0 otherwise), *CHRDS* (1 if at least one of 10 chronic conditions is present, 0 otherwise), *SMOK* (1 if presently smokes daily or if quit recently, 0 otherwise). All variables other than *LOSS* are measured at the start of the two-year period.

Note that the reference values for all variables included in the logistic model (2.1) are 0. The variables related to *BMI* (body mass index) originate from a variable *BMIGR* with three categories (0 for  $BMI \leq 18.5$ , 1 for  $18.5 < BMI < 25$ , and 2 for  $BMI \geq 25$ ). The ten chronic conditions considered were asthma, arthritis, back problems, bronchitis/emphysema, diabetes, heart disease, cancer, effects of stroke, urinary incontinence, and glaucoma/cataracts.

### 2.2 Simulated Finite Population

We simulated a finite population of 2.5 million individuals so that the individuals had some of the characteristics of the Canadian NPHS subpopulation of senior people, aged 65 and more and independent in the first cycle. The variables were generated as Bernoulli random variables using the joint probabilities estimated from the NPHS sample at Cycle 1.

After having simulated values for *SEX*, *AGEGR*, etc., the dependent variable *LOSS* was also created as a Bernoulli variable with probability equal to

$$p_x = p(\mathbf{LOSS} = 1 | \mathbf{x}) = \left[ 1 + \exp(-\mathbf{x}'\hat{\theta}) \right]^{-1},$$

where  $\mathbf{x}$  and  $\theta = (\beta_1, \beta_2, \dots, \beta_7)$  are defined by model (2.1). and where  $\theta$  was estimated from the NPHS sample to be  $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6) = (-3.799, 0.382, 1.388, 1.139, 0.406, 0.641, 0.484)$ . The proportion of individuals in the simulated finite population who lost their independence is 0.1009.

### 2.3 Simulated Clustering and Stratification of the Finite Population

We arranged the finite population into clusters in two different ways: one that is purely random (non-informative) clustering and the other which is, to some extent, an 'informative clustering.' In both cases the cluster sizes were between 20 and 60 individual records.

For the random clustering we randomly ordered the individual records and then assigned them to clusters whose sizes were generated as integers uniformly distributed between 20 and 60. For example, if the first random number was 29 the first 29 individual records went into the first cluster; if the second random number was 43, the next 43 individuals made up the second cluster, etc. In this way, 62,600 clusters were created, and the intracluster correlation was calculated to be 0.0001.

The informative clustering was done in the following way. From the first 1.875 million individual records, we created 62,600 clusters by the same random clustering procedure as explained above except that the cluster sizes were generated as random integers between 10 and 50. From the remaining 625 thousand records we created 62,500 groups of 10 records each, so that about 6260 groups had only records with LOSS =1, and the rest had only records with LOSS =0. Then, to each of the approximately, 6260 clusters with the largest proportion of records with LOSS=1 we added a group with ten ‘LOSS =1’ records. All other clusters received a group of ten records with LOSS =0. In this way, 62,600 clusters were created, with sizes between 20 and 60, and the intracluster correlation was 0.2637.

For each of the clusterings of the finite population, we arranged the clusters into strata in two different ways: (i) no stratification and (ii) two strata, where the first stratum contained the 25% of clusters having the largest proportion of records with ‘LOSS =1’, and the second cluster contained the remaining 75% of clusters.

## 2.4 Sample designs and generation of design information

We used a different sample design, depending on whether or not the population was stratified. In both cases, the sample design consisted of a sample of clusters chosen without replacement with the probability of selection proportional to the cluster size. The selection was done by the Sampford method as implemented in SAS procedure SURVEYSELECT. In the case of the unstratified population, we chose a sample of 30 clusters. In the case of the stratified population, we selected 15 clusters from each of the two strata, which meant that we were clearly oversampling from the first stratum, thus giving a larger probability of selection to clusters with larger  $p_i$  (i.e., larger proportion of records with LOSS=1).

The original sampling weights for the records included in a sample would be constant within clusters, since there was no subsampling within clusters. We then post-stratified these original weights to five poststrata based on known counts of (*AGEGR X SEX*) and *URBRUR*. After the poststratification, the weights of the records from the same cluster were not necessarily all equal..

For each sample of clusters we produced 500 bootstrap replicates. In the case of the unstratified design, for each bootstrap replicate we took a simple random sample with replacement of size  $n-1(=29)$  clusters. In the case of the stratified design, we selected a random sample with replacement of size  $n_h(=14)$  from the  $h$ -th stratum,  $h=1,2$ . The bootstrap weights in the  $b$ -th bootstrap replicate were then obtained by first adjusting the original sampling weights to reflect the inclusion (possibly more than once) of some clusters and the exclusion of other clusters, following the formula for the  $b$ -th bootstrap replicate:

$$w_{hij}^{(b)} = w_{hij} k_{hi}^{(b)} \frac{n_h}{n_h - 1},$$

where  $w_{hij}$  is the original sampling weight of the  $j$ -th individual from the  $i$ -th cluster in the  $h$ -th stratum,  $h=1,2$ , and  $k_{hi}^{(b)}$  is the number of repetitions of the  $i$ -th cluster in the  $b$ -th bootstrap replicate. Note that  $\sum_i k_{hi}^{(b)} = n_h - 1$ . These weights

were then poststratified to the known poststratum counts, in the same way as the original full-sample weights were calibrated.

## 2.5 Monte Carlo Setup

Considering the two types of clustering (non-informative and informative) and two stratification options (without and with), we had four different **population settings** with different levels of informativeness. We selected 500 Monte Carlo samples from each of the four population settings: (i) “Non-informative” clustering, no stratification, 30 clusters (**Low-Inf**); (ii) “Informative” clustering, no stratification, 30 clusters; (iii) “Non-informative” clustering, stratification, 15 clusters per stratum; and (iv) “Informative” clustering, stratification, 15 clusters per stratum (**High-Inf**). In this paper we present results for the two most extreme settings: the Low-Inf and the High-Inf settings.

## 2.6 Inferential approaches for logistic model

We considered the following five inferential approaches for fitting the logistic model to data from selected samples :

(i) **[DESIGN]** A full design-based approach where the variances are estimated by the linearized estimating equation bootstrap method (see Binder, Kovacevic, Roberts, 2004). The resulting estimates of the parameters and their variance estimates are denoted by  $\hat{\theta}_p$  and  $\hat{V}_p(\hat{\theta}_p)$ . The programming was done in SAS using SAS IML.

(ii) **[DESIGN-MODEL]** A combination of weighted point estimation of the parameters,  $\hat{\theta}_p$ , and model-based estimation of the variances using the robust sandwich variance estimator unadjusted for clustering,  $\hat{V}_\xi(\hat{\theta}_p)$ . This approach was implemented using the LOGISTIC procedure in SUDAAN and setting DESIGN=SRS, SEMETHOD=model and by specifying the survey weight variable in a WEIGHT statement.

(iii) **[MODEL]** A model-based approach (unweighted) where both the point estimates  $\hat{\theta}_\xi$  and their variances  $\hat{V}_\xi(\hat{\theta}_\xi)$  are calculated as if the sample design was simple random sampling of individuals with replacement. This approach was implemented using the LOGISTIC procedure in SUDAAN and specifying DESIGN=WR, SEMETHOD=model and without using the WEIGHT statement.

(iv) **[NLMIXED]** A model-based approach where the effects of clustering  $u_j$  are modeled as additive random effects:

$$\xi_1 : \text{logit}(y_{ij}) = x'_{ij} \theta + u_j + \varepsilon_{ij}.$$

The model parameters  $\theta$  are estimated by a model-based (unweighted) estimate  $\hat{\theta}_{\xi_1}$  and the corresponding variance matrix is estimated by the model-based variance estimator  $\hat{V}_{\xi_1}(\hat{\theta}_{\xi_1})$ . The estimation was done using SAS PROC NLMIXED. Note that the model did not account for the stratification in the survey design.

(v) **[NLMIXED-WGT]** Fitting the  $\xi_1$  model as in (iv) but using the survey weights, obtaining the point estimates  $\hat{\theta}_w$  and their model-based variance matrix using the “sandwich” estimator evaluated at the final estimate values,  $\hat{V}_{\xi_1}(\hat{\theta}_w)$ .

Note that in (iv) and (v) we are not interested in estimating the variance components. The estimation was done using SAS PROC NLMIXED with the weight variable specified in the REPLICATE statement. (Since the variable specified in the REPLICATE statement must have integer values, and since the survey weights had 4 digits after the decimal point, the variable specified in the REPLICATE statement was actually 10,000\*weight.)

## 3. IMPACT OF INFORMATIVENESS ON POINT AND VARIANCE ESTIMATES

We considered a variety of measures for comparing the different approaches. In the descriptions of the measures below we use the subscript M to denote any one of the approaches and we use  $\theta$  to represent any one of the coefficients in our model.

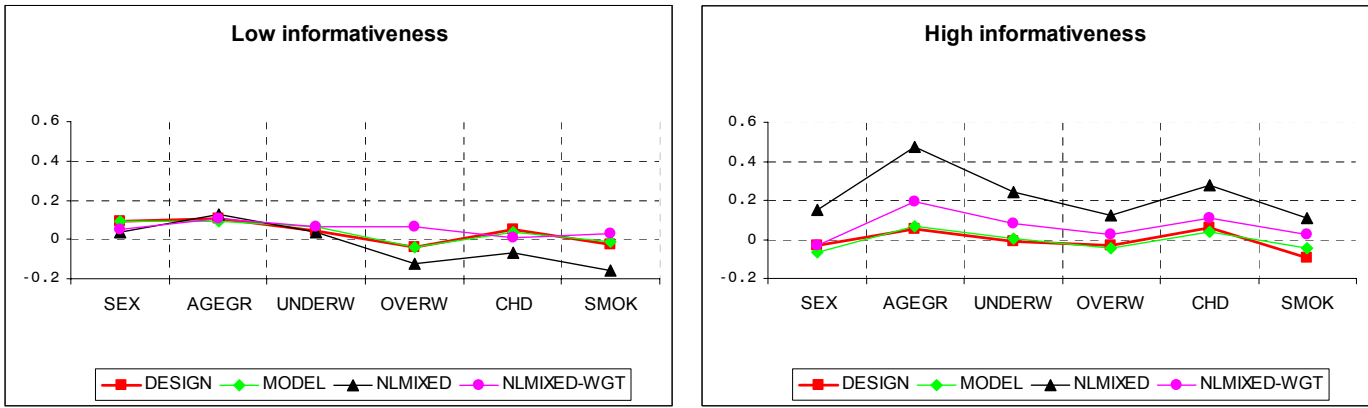
The first measure that we look at, which examines the impact of informativeness on point estimates, is the **standardized difference of a parameter estimate from the true value**:

$$\frac{E_{sim}(\hat{\theta}_M) - \theta}{\sqrt{MSE_{sim}(\hat{\theta}_M)}}, \text{ where} \quad (3.1)$$

$E_{sim}(\hat{\theta}_M) = \frac{1}{500} \sum_k \hat{\theta}_{M,k}$ ,  $MSE_{sim}(\hat{\theta}_M) = \frac{1}{500} \sum_k (\hat{\theta}_{M,k} - \theta)^2$ , and  $\hat{\theta}_{M,k}$  is the estimate of  $\theta$  using the M-th approach with

the k-th Monte Carlo sample. The closer this measure is to 0, the better is the parameter estimate. We show our results in Chart 1 for the 6 model coefficients other than the intercept. The left-hand graph shows the results for the Low-Inf case and the right-hand graph is for the High-Inf setting. The variables from the model are given along the horizontal axis, and the magnitude of the measure is given along the vertical axis. Values of the measure for the different variables for the same approach are connected by lines of particular colors.

**Chart 1.** Standardized difference of a parameter estimate from the true value



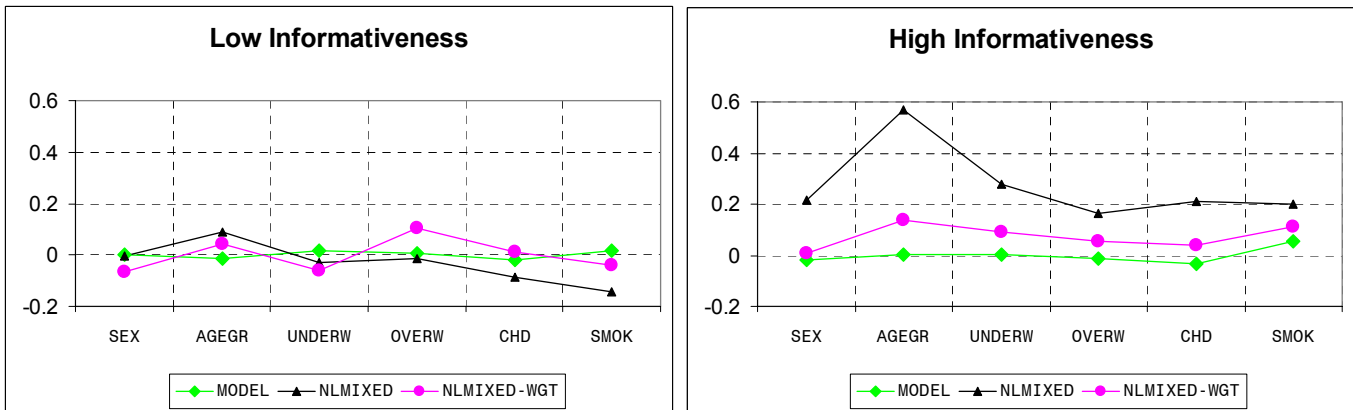
For the Low-Inf setting all approaches perform similarly, with the measure staying close to 0 for all variables; a slight exception is the NLMIXED approach. For the High-Inf setting, the plots are more scattered for the different approaches; however, the DESIGN and MODEL approaches yield very similar results and are the closest to the zero line. The DESIGN-MODEL approach yields exactly the same results as the DESIGN approach, and thus is not shown on these graphs. In a real survey situation, since we do not know the true value of  $\theta$ , we cannot calculate this statistic.

A measure which can be calculated from survey data is **the standardized difference between an alternative parameter estimate  $\hat{\theta}_M$  and the design-based parameter estimate  $\hat{\theta}_p$** . This measure, which is an average over the 500 Monte Carlo samples of the standardized difference, is defined as follows:

$$E_{sim} \left\{ \frac{\hat{\theta}_M - \hat{\theta}_p}{\sqrt{\hat{V}_p(\hat{\theta}_p)}} \right\} = \frac{1}{500} \sum_k \frac{\hat{\theta}_{M,k} - \hat{\theta}_{p,k}}{\hat{V}_p(\hat{\theta}_{p,k})}. \tag{3.2}$$

The results are given in Chart 2.

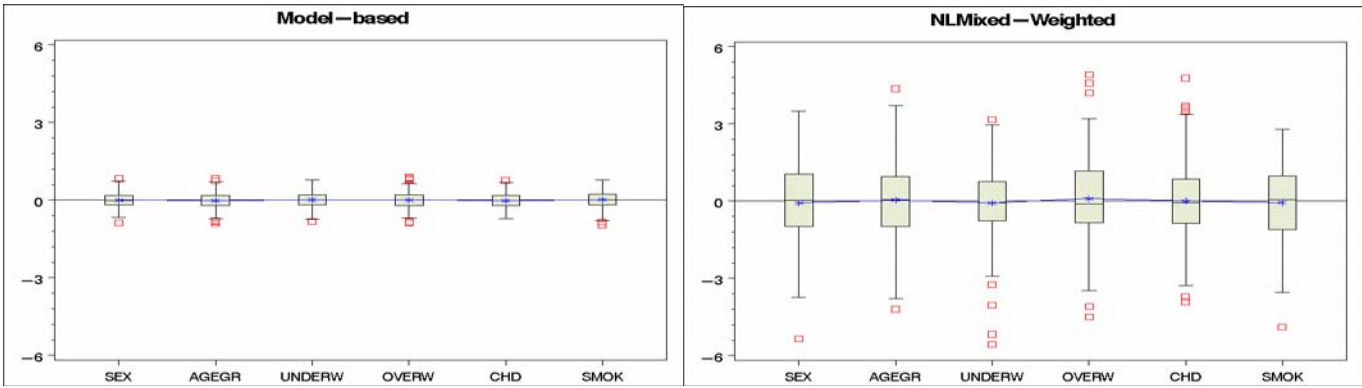
**Chart 2:** Standardized difference between an estimate  $\hat{\theta}_M$  and the design-based estimate  $\hat{\theta}_p$  (average over 500 samples)



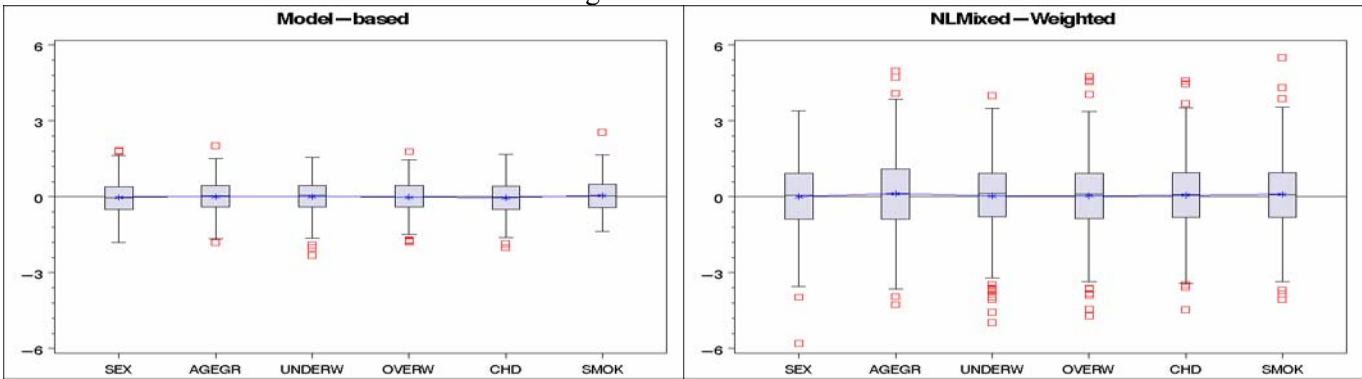
In the Low-Inf case, all the approaches yield lines close to 0, with the MODEL approach staying the closest. In the Hi-Inf case there is a larger spread in the plots, although the model-based is still close to zero. The NLMIXED shows the most extreme behaviour. Note that, again, there is no line for the DESIGN-MODEL approach since the parameter estimates under this approach are the same as the DESIGN approach. Since this measure is calculable from a single sample we want to consider it as a possibility for identifying informativeness.

The standardized difference measure defined above and illustrated in Chart 2 is an average of a standardized difference over 500 Monte Carlo samples. It would also be of interest to see whether the variability among the Monte Carlo samples in the size of the standardized difference changes with the estimation approach. Chart 3 contains the box-and-whisker plots for the MODEL and the NLMIXED-WGT approaches since they exhibited a similar average behaviour, as seen in Chart 2.

**Chart 3.** Box-and-whisker plots of 500 values of measure (3.2) for the MODEL and NLMIXED-WGT approaches  
Low Informativeness



High Informativeness

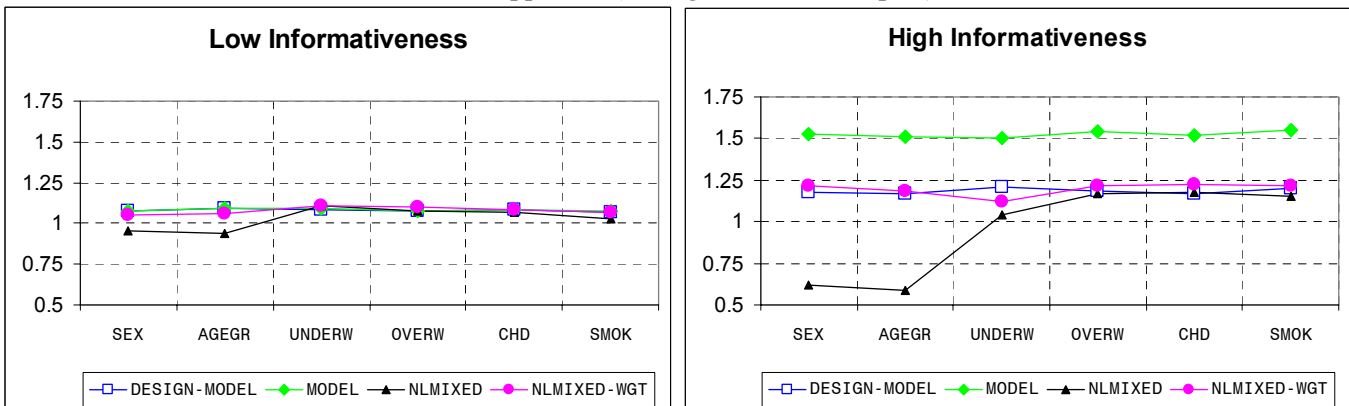


We see that the NLMIXED-WGT approach is much more variable than the MODEL approach under both the Low-Inf and the Hi-Inf setups, and that Hi-Inf generally yields a larger scatter of differences than Low-Inf.

The previous two measures compared the point estimates. The next one compares variances by computing **the ratio of variance**  $\hat{V}_p(\hat{\theta}_p)$  from the design-based approach to the variance  $\hat{V}_M(\hat{\theta}_M)$  from an alternative approach. In Chart 4 we present the averages of these ratios over the 500 Monte Carlo samples:

$$E_{sim} \left\{ \frac{\hat{V}_p(\hat{\theta}_p)}{\hat{V}_M(\hat{\theta}_M)} \right\}. \quad (3.3)$$

**Chart 4:** The ratio of variance  $\hat{V}_p(\hat{\theta}_p)$  from the design-based approach to the variance  $\hat{V}_M(\hat{\theta}_M)$  from an alternative approach (average over 500 samples)



In this chart the degree of closeness of an alternative approach to the design-based approach with respect to variance estimation is indicated by how close the measure is to the value 1. In the Low-Inf case, all approaches were very similar and all had measure values close to 1. In the Hi-Inf case, we see more spread in plots and none of the approaches gives values of 1. The MODEL approach is the farthest away from the design-based for this particular measure, while, for the previous two measures we couldn't distinguish between these two approaches. These results indicate that we need some comparison of variances when assessing informativeness. We will address the issue of choice of measure in more depth in Section 5.

#### 4. IMPACT OF INFORMATIVENESS ON POWER (AND SIZE)

We now consider the impact of informativeness on hypothesis testing about the model coefficients. We assess the impact of informativeness on the power (and size) of tests by considering two examples: in the first example the null hypothesis being tested is actually true in the population by construction, and in the second example, the null hypothesis is false in the population by construction.

**Example 1:** The initial model augmented by a SEX×AGEGR term is fit to sample data sets. The null hypothesis is then formulated as “ $H_0$ : There is no interaction between SEX and AGEGR.” The test statistic generally used for this hypothesis is the Wald statistic  $\hat{X}_W^2 = \hat{\theta}_{Sex \times Agegr}^2 / \hat{V}(\hat{\theta}_{Sex \times Agegr})$  which is then compared to the 95<sup>th</sup> percentile of a  $\chi_1^2$  distribution ( $\chi_1^2(.95) = 3.841$ ).

We developed a strategy for deriving power curves  $\gamma(\theta_a)$  for testing the hypothesis that a coefficient is equal to zero. In this strategy, the Wald statistic for approach M, denoted by  $\hat{X}_{W(M)}^2$ , is assumed to be proportional to a non-central  $\chi^2$  variable with one degree of freedom so that the power curve for approach M is

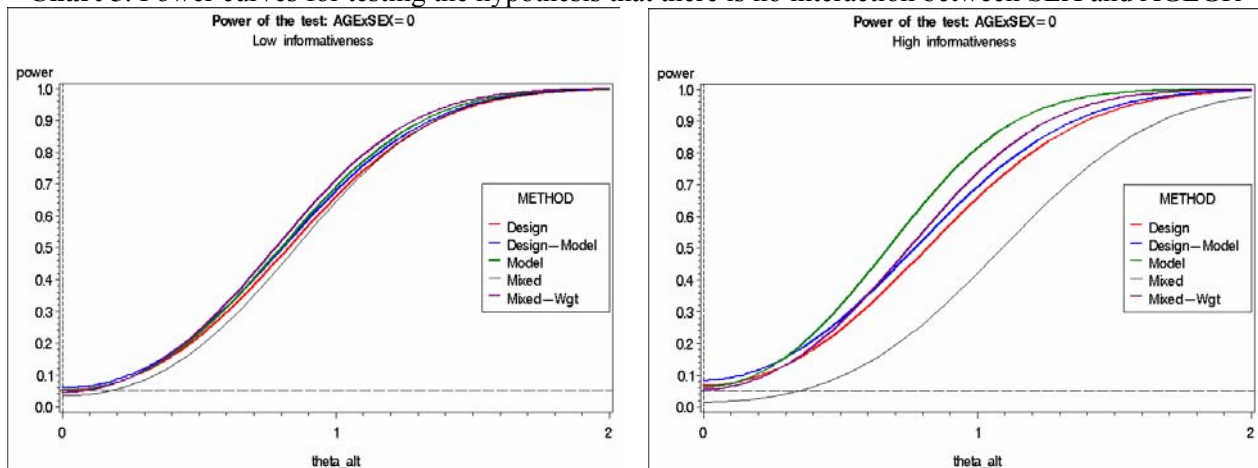
$$\gamma_M(\theta_a) = Prob\{\hat{X}_{W(M)}^2 > 3.841 \mid \theta = \theta_a, \hat{X}_{W(M)}^2 \sim c_M \chi_1^2(b_{a(M)})\},$$

where the non-centrality parameter is  $b_{a(M)} = d'_{a(M)} V_{(M)}^{-1} d_{a(M)}$  with  $d_{a(M)} = \theta_a + bias_p(\hat{\theta}_M) = E_p(\hat{\theta}_M \mid \theta = \theta_a)$ , and the proportionality parameter is  $c_M = V_M^{-1} V_{(M)}$ , where  $V_M = E_p(\hat{V}_M(\hat{\theta}_M))$  and  $V_{(M)} = V_p(\hat{\theta}_M)$ . For the simulated power curves, both of these variances are estimated by taking averages of appropriate estimates over 500 samples as shown in Table 1. A new approximation was developed for the non-central case for more than 1 df and will be reported elsewhere.

**Table 1. Variance approximations for power calculations**

	Approach <i>M</i>				
	DESIGN	DESIGN-MODEL	MODEL	NLMIXED	NLMIXED-WGT
$V_M$	$E_{sim}(\hat{V}_p(\hat{\theta}_p))$	$E_{sim}(\hat{V}_\xi(\hat{\theta}_p))$	$E_{sim}(\hat{V}_\xi(\hat{\theta}_\xi))$	$E_{sim}(\hat{V}_{\xi_1}(\hat{\theta}_{\xi_1}))$	$E_{sim}(\hat{V}_{\xi_1}(\hat{\theta}_w))$
$V_{(M)}$	$V_{sim}(\hat{\theta}_p)$	$V_{sim}(\hat{\theta}_p)$	$V_{sim}(\hat{\theta}_\xi)$	$V_{sim}(\hat{\theta}_{\xi_1})$	$V_{sim}(\hat{\theta}_w)$

**Chart 5. Power curves for testing the hypothesis that there is no interaction between SEX and AGEGR**



In Chart 5, we give just half of each power curve since the curves are symmetric around 0. The low-informativeness power curves are very similar across the methods. The high-informativeness power curves spread farther apart with the NLMIXED most separated from the others.

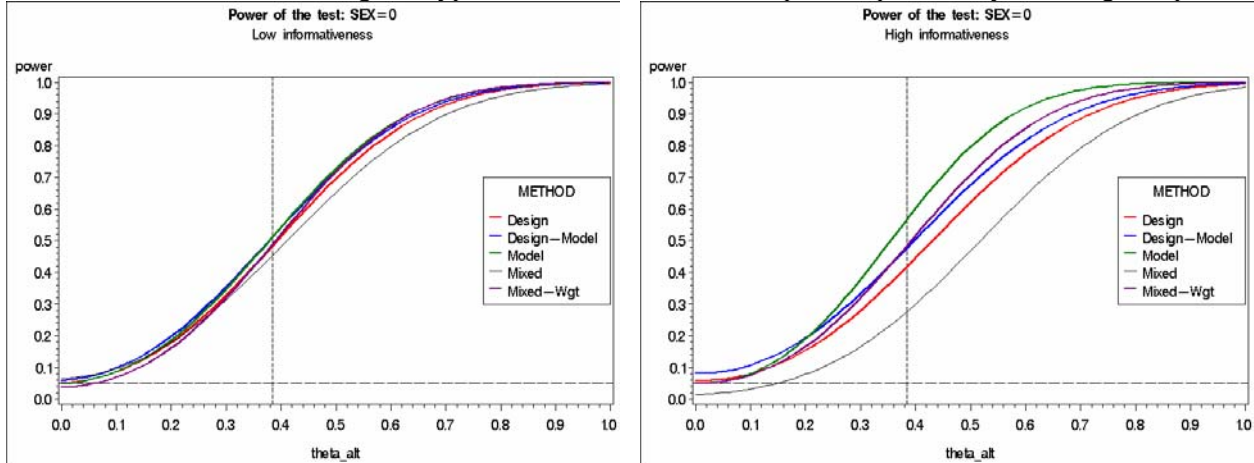
Table 2 shows the value of the proportionality coefficient  $c$ , the value of the theoretical power curve at the true alternative (which is 0) and the empirical rejection rate when the hypothesis was tested using the 500 Monte Carlo samples. The power and rejection rate values are quite close, regardless of approach, indicating that the power curve at 0 is well approximated by the empirical rejection rate. The power and the rejection rates are closer to the nominal 5% level for the Low-Inf case than for High-Inf. The fact that the values for the DESIGN approach are above the nominal 5% level might be attributed to the small sample sizes. The NLMIXED approach seems to be performing differently from the other approaches.

**Table 2.** Proportionality coefficient, theoretical power at  $\theta_a=0$ , and the empirical rejection rates

Informativeness	Low			High			
	Approach $M$	$c = V_M^{-1}V_{(M)}$	Theoretical power at $\theta_a=0$ (size)	Empirical Rejection Rate	$c = V_M^{-1}V_{(M)}$	Theoretical power at $\theta_a=0$ (size)	Empirical Rejection Rate
DESIGN		<b>1.029</b>	<b>0.053</b>	<b>0.058</b>	<b>1.143</b>	<b>0.067</b>	<b>0.080</b>
DESIGN-MODEL		<b>1.084</b>	<b>0.060</b>	<b>0.051</b>	<b>1.277</b>	<b>0.083</b>	<b>0.082</b>
MODEL		<b>0.999</b>	<b>0.050</b>	<b>0.053</b>	<b>1.102</b>	<b>0.063</b>	<b>0.058</b>
NLMIXED		<b>0.870</b>	<b>0.036</b>	<b>0.014</b>	<b>0.657</b>	<b>0.016</b>	<b>0.014</b>
NLMIXED-WGT		<b>0.948</b>	<b>0.044</b>	<b>0.049</b>	<b>1.038</b>	<b>0.054</b>	<b>0.060</b>

**Example 2:** Consider a case where the null hypothesis is not true in the population. In particular, consider the hypothesis “ $H_0$ : SEX has no impact on the probability of losing independence.” The coefficient on the SEX variable is actually equal to 0.384 and this value is indicated by a vertical line on Chart 6. The order and spread of the power curves remain quite similar to those seen in the previous example; however, this means that, for the Hi-Inf case, there is a much wider range in the values of the curves at the true value. Nonetheless, as can be seen in Table 3, the theoretical power and the empirical rejection rates are quite close for both Hi-Inf and Low-Inf.

**Chart 6.** Power curves for testing the hypothesis that SEX has no impact on probability of losing independence



**Table 3.** Proportionality coefficient, theoretical power at  $\theta_a=0.384$  and the empirical rejection rates

Informativeness	Low			High			
	Approach $M$	$c = V_M^{-1}V_{(M)}$	Theoretical power at $\theta_a=.384$	Empirical Rejection Rate	$c = V_M^{-1}V_{(M)}$	Theoretical power at $\theta_a=.384$	Empirical Rejection Rate
DESIGN		<b>0.989</b>	<b>0.483</b>	<b>0.505</b>	<b>1.070</b>	<b>0.416</b>	<b>0.450</b>
DESIGN-MODEL		<b>1.061</b>	<b>0.511</b>	<b>0.510</b>	<b>1.261</b>	<b>0.475</b>	<b>0.484</b>
MODEL		<b>0.971</b>	<b>0.510</b>	<b>0.527</b>	<b>1.006</b>	<b>0.566</b>	<b>0.584</b>
NLMIXED		<b>1.131</b>	<b>0.453</b>	<b>0.448</b>	<b>0.636</b>	<b>0.275</b>	<b>0.246</b>
NLMIXED-WGT		<b>0.889</b>	<b>0.488</b>	<b>0.480</b>	<b>0.997</b>	<b>0.483</b>	<b>0.485</b>

## 5. PROPOSED MEASURE OF INFORMATIVENESS

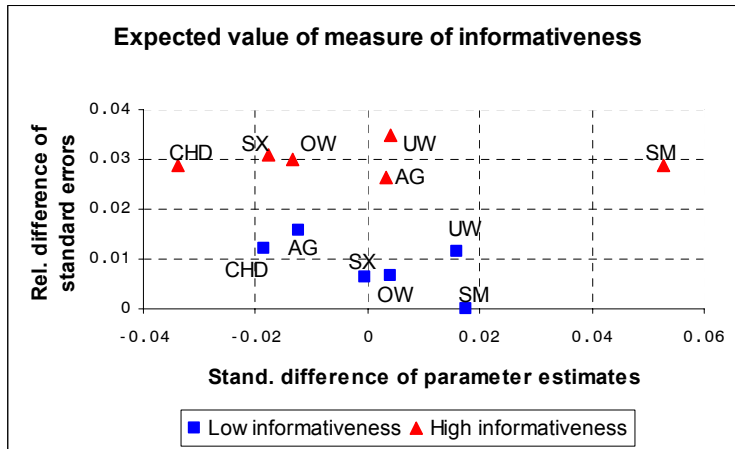
We are presuming that the researcher wishes to have as his basis for statistical inference the distribution of the observations resulting from generating a finite population under a model and then selecting the sample using the sampling design. Since the true model generating the finite population was the one used in the MODEL approach, we are going to restrict our attention to that model in this section. If the sample design is not informative, we know that  $E_{\hat{\varphi}_p}(\hat{\theta}_p) = E_{\hat{\varphi}_p}(\hat{\theta}_\xi)$  and  $E_{\hat{\varphi}_p}(\hat{V}_\xi(\hat{\theta}_p)) = E_{\hat{\varphi}_p}(\hat{V}_p(\hat{\theta}_p))$  (Binder and Roberts, 2003). Based on this and on what we have observed in previous sections, we propose a bivariate measure for assessing informativeness: one component should compare the point estimates and the other should deal with the variance estimation. We thus suggest the following measure – which can be calculated from a single sample - for comparing the DESIGN and MODEL approaches:

$$\left\{ \frac{\hat{\theta}_\xi - \hat{\theta}_p}{\sqrt{\hat{V}_p(\hat{\theta}_p)}}, 1 - \sqrt{\frac{\hat{V}_\xi(\hat{\theta}_p)}{\hat{V}_p(\hat{\theta}_p)}} \right\}.$$

The particular transformation of the variance estimates shown here has been chosen to ensure that it is of the same order of magnitude as the first component. Both components would be close to 0 in the situation of non-informativeness.

We have created a scatter plot of the simulation average versions of this measure in Chart 7:

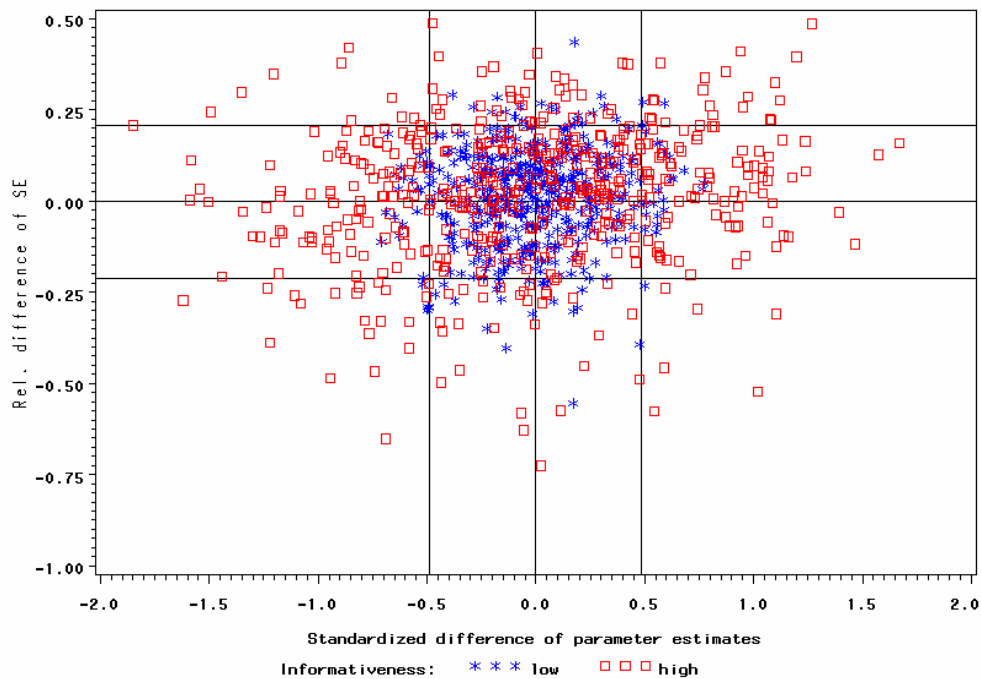
**Chart 7:** Average values of proposed measure of informativeness over 500 samples



There is a clear separation between the Low-Inf and High-Inf cases in the average variance values of the bivariate measure which are plotted on the vertical axis, while any distinction regarding the average point estimate values is less obvious except for the larger spread for the high informativeness case.

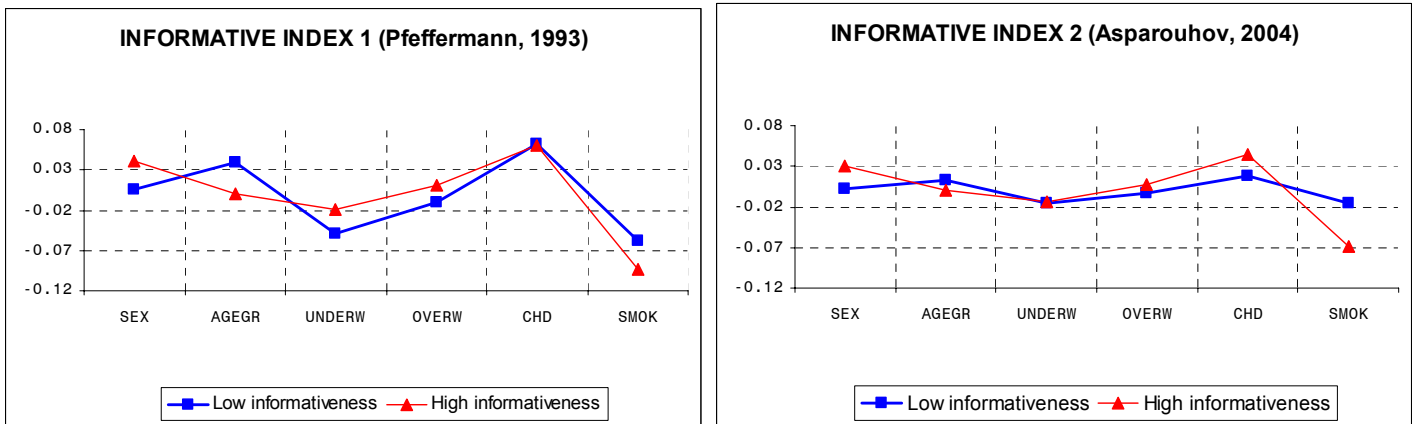
In Chart 8, we use the individual results from the 500 Monte Carlo samples under the Low-Inf and High-Inf cases to display a scatter plot of the proposed measure for the coefficient on CHD, the variable with the least difference in point estimates between the MODEL and DESIGN approaches. On both axes we show the 5<sup>th</sup> and the 95<sup>th</sup> percentiles of the points from the Low-Inf case. While a very high percentage of the Low-Inf points are within the square, only 50% of the High-Inf points are there.

**Chart 8.** Scatter plot of the proposed measure obtained over 500 samples for the variable CHD for two level of informativeness.



It should be noted that there are other measures of informativeness proposed in the literature (e.g., Dumouchel and Duncan (1983), Pfeffermann (1993), and Asparouhov (2004)). However, all of them are univariate and only compare point estimates. In the following chart we present the averages over 500 samples of the Pfeffermann Index and the Asparouhov Index. Obviously, these two measures are not distinguishing the low and high informativeness cases when the model-based and design-based point estimates are very similar.

**Chart 9.** Pfeffermann’s (1993) Informative Index and Asparouhov’s (2004) Informative Index averaged over 500 samples



## 6. CONCLUDING REMARKS

In this paper we study the impact of informativeness on some of the substantive conclusions from an analysis of survey data. We generated a finite population and developed a tool for varying the level of informativeness of the sample design in order to compare the performance of several different analytical approaches under different levels of informativeness. We found that, for every approach considered, there were differences in the results between the low informativeness and the high informativeness cases, and that approaches differed. Thus, informativeness – and how it is handled - does matter.

We included two simple mixed model approaches having a random component that was intended to account for the clustering in the survey designs. In general, the mixed model did not perform well, indicating that it is not sufficient to account for the sample design simply by including a random cluster effect in the model as is often recommended in analytic papers. We found that using weighted estimation in the mixed model improved results slightly. We also noted that the design-based approach was not clearly best across all measures used in this simulation study. One of the reasons for this could be the small sample sizes used in this simulation study. We anticipate that the performance of the design-based approach would considerably improve with an increase of sample sizes or an increase in the complexity of the sampling design. We would like to study this further. For assessing informativeness we suggested that a bivariate approach is required but we need more research before we can propose a formal bivariate test for significant informativeness of the sample design.

## REFERENCES

- Asparouhov, T. (2004). Weighting for Unequal Probability of Selection in Multilevel Modeling. *Mplus Web Notes*: No.8.
- Binder, D.A., Kovacevic, M.S., and Roberts, G. (2004). Design-Based Methods For Survey Data: Alternative Uses Of Estimating Functions. *Proceedings Of The Section On Survey Research Methods*, 3301-3312, JSM, Toronto
- Binder, David A. and Georgia R. Roberts (2001). Can informative designs be ignorable? *Newsletter of the Survey Research Methods Section, Issue 12*, American Statistical Association.
- Binder, David A. and Georgia R. Roberts (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, 29-48.
- DuMouchel, W.H. and Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. *Journal of American Statistical Association*, 78, 535-543.
- Martel, L., Bélanger, A., and Berthelot, J.-M. (2002), "Loss and Recovery of Independence among Seniors," *Health Reports*, 13, 35-48.
- Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review*, 317-338.
- Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992) Some Recent Work On Resampling Methods For Complex Surveys. *Survey Methodology*, 18, 209-217.