Sample Design for Longitudinal Multiphase Samples with Misclassification

Andrea Piesse[1] and Sharon Lohr[1]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850

Abstract
Surveys concerned with substance use or the etiology of medical conditions often want to have larger samples of persons who are in specific subpopulations, while maintaining the ability to monitor transitions from one subpopulation to another. We consider optimal survey design using a two-phase sample, in which persons are selected for the first phase using a screening instrument that misclassifies some individuals, and the second-phase sample is selected using a more accurate classification instrument. We discuss software that can be used to solve the optimization problem with constraints, and illustrate with an example.

Key Words: constrained optimization, optimal design

1. Introduction

Studies of social and medical trends over time often entail obtaining a representative sample of persons in specific subpopulations. The National Longitudinal Study of Adolescent to Adult Health (Add Health), for example, follows a sample of adolescents over time in order to study social mechanisms associated with health and high-risk behaviors (Boonstra, 2001). This was a two-phase survey in which students from sampled schools answered a brief questionnaire in the first phase; in the second phase, after parental consent was obtained, more extensive interviews on a subsample were taken at the students' homes on topics such as sexual activity and substance use. The original interviews were conducted during the 1994-1995 school year, and follow-up interviews were conducted in 1996, 2001-2001, and 2008-2009, with Wave V data collection planned for 2016-2018 (Harris, 2007; Harris et al., 2015; Carolina Population Center, 2015).

For some surveys it is desired to have larger samples of persons who are in specific domains of interest (subpopulations), while maintaining the ability to monitor transitions from one domain to another. In Add Health, while the core sample was essentially self-weighting, supplemental samples were taken of students in specific domains based on race/ethnicity, multiple birth status, adoption status, disability, and parental education (Harris, 2007). The National Survey on Drug Use and Health (NSDUH; Morton et al., 2013) used higher sampling rates for persons ages 12 to 25 and persons in smaller states. A longitudinal study in psychiatry may oversample persons who meet clinical standards for depression in the baseline interview. A survey intended to study diabetes risks and management may oversample persons with diagnosed diabetes, undiagnosed diabetes, and pre-diabetes.

In these studies, persons to have a higher probability of selection cannot be identified from the sampling frame, and an accurate classification may be expensive. To reduce costs, a two-phase sampling procedure may be used, in which an inexpensive screener instrument provides a preliminary classification and an instrument used in the second phase gives a more accurate classification.

In a household survey, one household member might provide the initial classification for all adults in the household. The household respondent might not know the substance use of other household members. Often, an initial screener instrument is briefer than the follow-up instrument, so the household respondent might initially provide an incorrect classification for his or her own substance use—for example, a respondent might report that he does not use marijuana in the screener interview but the answers to more detailed questions might classify him as a marijuana user. Context effects associated with the setting of the first-phase screener might also lead to misclassification through reporting of responses that are considered to be more socially desirable.

In a psychiatric study, a screener instrument may give an initial classification but an in-depth interview or test may result in a different classification. Persons with undiagnosed diabetes or pre-diabetes are often unaware that they have the condition, so some individuals are misclassified if a questionnaire is used to screen participants. A follow-up, more accurate screening would use blood glucose or HbA1c to determine diabetes status.

The focus in this paper is on surveys that have the following common features:

- The population can be separated into mutually exclusive sampling strata whose union is the entire population.
- Information on sampling stratum membership is not present on the sampling frame.
- In some cases, there are no accurate estimates of the population percentage in each sampling stratum before the survey is fielded.
- An initial screening instrument provides sampling stratum membership but misclassifies some individuals. This instrument is used to determine which persons are selected in the first-phase sample.
- The persons selected in the first-phase sample are administered an instrument that provides a more accurate classification. The first- and second-phase classifications are used in the second phase of sampling to select persons for participation in the study.
- Domains of interest are combinations of mutually exclusive sampling strata.
- It is desired to achieve predetermined precision levels (and possibly minimum sample sizes) for specific domains of interest as well as for the population as a whole. This requires higher sampling fractions in some sampling strata.
- Individuals may be followed over time and may transition to different domains in later interviews. If persons who are selected in the second phase of sampling at time 1 are periodically reinterviewed over time, it may be desired to achieve predetermined precision levels for individuals in specific domains later in the study.

The disparate goals require compromises in the sample allocation. Proportional allocation would be most efficient for estimates pertaining to the whole population, but would not give the precision needed in small domains. On the other hand, disparate sampling

fractions lead to weight variation which increases the variance for estimates (Kish, 1992; Kalton et al., 2005). The misclassification may also lead to weight variation within the domains of interest: if persons who are unaware they have diabetes are sampled at a much lower rate in the first phase, even if they are sampled with certainty in the second phase they may still have much higher weights than persons who are accurately classified as having diabetes in the first phase. If design parameters such as the population percentages in each sampling stratum or the misclassification probabilities are unknown before the survey commences, then the procedure used to assign selection probabilities in the first phase and second phase must be flexible in order to allow the design goals to be met.

In this paper, we describe nonlinear programming methods that may be used to optimize the first- and second-phase sample sizes in different sampling strata in order to meet multiple objectives and constraints. Section 2 describes constraints that may be used for longitudinal surveys with misclassification, and formalizes the goals as an optimization problem. It also describes steps taken to ensure that the design is flexible and can be modified to adapt to updated estimates of the misclassification probabilities and population sizes for the first-phase sampling strata. Section 3 describes software that may be used to optimize the survey design and presents an example for a hypothetical survey using Microsoft Excel® and SAS® software, and Section 4 discusses the results.

2. Goals and Constraints for the Sampling Design

2.1 Information Supplied by User

The overall goal of any sampling design is to collect information on the population of interest as efficiently as possible. For the types of surveys considered in this paper, the user needs to provide the following information:

- Mutually exclusive first-phase sampling strata that together comprise the entire population. In a survey on diabetes, these might be cross classifications of age (under 50, or 50+), race/ethnicity (Native American or Alaska Native, Hispanic, white non-Hispanic, African American non-Hispanic, other), and diabetes status (diagnosed diabetes, diagnosed pre-diabetes, undiagnosed diabetes, undiagnosed pre-diabetes, no diabetes). Individuals are assigned to these strata based on the classification from the first-phase screener instrument.
- Mutually exclusive second-phase classifications, from the second-phase instrument that has more accurate information. For example, these might be cross-classifications of the same age, race/ethnicity, and diabetes status grouping used at the first phase, but are based on the information from the second-phase instrument.
- Estimates of population sizes for the first-phase sampling strata and of the misclassification probabilities. Denote the estimated probability of $P($second-phase classification is $k \mid$ first-phase stratum is $i)$ by p_{ik}.
- Anticipated response rates to the second phase for persons selected at the first phase.
- Composition of the domains of interest, which can overlap, based on the second-phase classification. In a survey on trajectories of diabetes, domains might include persons under age 50, persons ages 50 and over, African Americans, Hispanics, Native Americans or Alaska Natives, Hispanics under age 50, persons with diabetes, persons under age 50 with pre-diabetes, and the entire population.

We include the entire population as one of the domains to ensure that all population members are assigned a nonzero inclusion probability.

- Desired minimum precision (and possibly minimum sample size) in each domain of interest.
- An objective function to be minimized. This might be a weighted average of the estimated variances for the domains, or, if the variances are specified, a measure of total cost. In the example presented in this paper, we use a weighted average of standard errors for the domains.
- Cost to obtain (1) first-phase screening information, (2) a second-phase screening interview/test to obtain correct classification, (3) an extended survey interview or medical examination information about the person selected at the second phase, and (4) follow-up interviews at later times. These costs should include the costs associated with nonresponse.
- If desired, upper bounds on the coefficient of variation due to weights or on the ratio (maximum weight)/(minimum weight) within or across different second-phase classifications.
- (If a longitudinal study) estimates of transition probabilities among sampling strata over time, and desired minimum precisions for domains at later points in time.

In addition, structural constraints need to be considered, such as specifying that all probabilities for sample inclusion are between 0 and 1. These constraints are defined mathematically in subsection 2.2.

Many of the parameters specified for the design, such as the estimated population sizes for the first-phase sampling strata, the anticipated response rates, and the estimated misclassification probabilities, may be very rough estimates. As information accrues from early data collection in the survey, these estimates can be updated to inform later data collection.

It is common practice in surveys to collect data in release groups. For a survey that is to be conducted over a one-year period, for example, 1/6 of the sample might be released every second month over the survey period. The optimization can be set up so that information about misclassification, response propensity, and first-phase stratum membership found from the early release groups can be used to refine the selection probabilities for later release groups.

2.2 Nonlinear Constrained Optimization Problem

The goals and constraints for the sample design can be formalized as an optimization problem that involves a nonlinear optimization function as well as nonlinear constraints. Let K = the number of first-phase sampling strata comprising the population. With the more accurate classification using the second-phase instrument, a person in first-phase sampling stratum k may be determined to actually have that classification, or he/she may be determined to belong to one of the other $K-1$ categories. Consequently, when selecting persons at the second phase for the extended interview, sampling rates need to be determined for each combination of first-phase preliminary classification and second-phase actual classification, resulting in up to K^2 possible second-phase sampling strata. The parameters to be optimized are the sample sizes $\boldsymbol{\theta} = [\mathbf{n}'\ \mathbf{m}']'$, where $\mathbf{n} = [n_1, n_2, \ldots, n_K]'$ is the K-vector of first-phase sample sizes and $\mathbf{m} = [m_{11}, m_{12}, m_{1K}, m_{21}, \ldots, m_{KK}]'$ is the K^2-vector of second-phase sample sizes, with m_{ik} being the

second-phase sample size for persons in first-phase stratum i with second-phase classification k. The sample size m_{ik} depends on the estimated misclassification probabilities, with $m_{ik} = n_i r_{1i}\, p_{ik} a_{ik}$, where r_{1i} is the anticipated response rate to the second-phase instrument and a_{ik} is the subsampling rate in the second phase. With this formulation, the second-phase sample size of persons having second-phase classification k equals $\sum_i m_{ik}$.

The Q domains of interest, indexed from $q = 1, \ldots, Q$, are formed as combinations of the K mutually exclusive second-phase classifications that together form the population. It is assumed that everyone in the population is in at least one of the Q domains. This is typically achieved by having one of the domains consist of the entire population.

We assume that a simple random subsample is taken of persons within the K first-phase and K^2 second-phase sampling strata. Variance inflation arises because of weight variation. The variance of an item with population proportion p among persons with second-phase classification k is

$$V(\hat{p}) = \frac{p(1-p)}{\sum_i m_{ik}} \text{weff}_k \qquad (1)$$

where weff_k is the design effect due to weight variation in second-phase classification k: using the formula in Kish (1992),

$$\text{weff}_k = \frac{\sum_i m_{ik} \sum_i m_{ik} w_{ik}^2}{(\sum_i m_{ik} w_{ik})^2}, \qquad (2)$$

$$w_{ik} = \frac{1}{P(\text{select person in } (ik) \text{ for sample})} = \frac{e_i}{n_i} \frac{n_i r_{1i} p_{ik}}{m_{ik}} = \frac{e_i r_{1i} p_{ik}}{m_{ik}}, \qquad (3)$$

e_i is the estimated number of screened persons in first-phase stratum i, and r_{1i} is the anticipated second-phase response rate for persons sampled in first-phase stratum i. For some surveys, the extended interview may immediately follow the second-phase screening and it could be assumed that all persons selected will complete the interview. In other surveys, the extended interview or test may be separate or burdensome, and there may be additional nonresponse for persons selected at the second phase. If desired, equations (1), (2), and (3) may be modified to include a response rate r_{2ik} to account for the anticipated interview response rate for persons sampled in first-phase stratum i with second-phase classification k. The domain formulas below would be modified similarly.

For domain q, we have

$$V_q(\hat{p}) = \frac{p(1-p)}{\sum_{k \in q} \sum_i m_{ik}} \text{weff}_q,$$

where

$$\text{weff}_q = \frac{(\sum_{k \in q} \sum_i m_{ik})(\sum_{k \in q} \sum_i m_{ik}\, w_{ik}^2)}{\left(\sum_{k \in q} \sum_i m_{ik}\, w_{ik}\right)^2}.$$

The following constraints are used for the optimization.

1. $1 \leq n_i \leq e_i$.
2. If $p_{ik} = 0$ then $m_{ik} = 0$.
3. For each value of k for which $p_{ik} > 0$, $1 \leq m_{ik} \leq n_i r_{1i} p_{ik}$.
4. $\sum_i m_{ik} \geq M_k$. Each domain achieves a minimum sample size M_k which is specified by the user.
5. Total cost $= c_1 \sum_i n_i + c_2 \sum_i \sum_k m_{ik} \leq C$, where C is the total budget for interviewing at time 1.
6. Each domain estimate achieves a predetermined precision. This can be expressed in terms of the domain variance or domain standard error. In our example, we constrained $SE_q(\hat{p}) \leq S_q$ for $q = 1, \ldots, Q$.
7. In a longitudinal survey, some persons may change domain status over time. For example, some persons without diabetes may transition to pre-diabetes status. Consequently, the set of persons with pre-diabetes at time 2 will consist of persons having different second-phase classifications and sampling weights at time 1. This leads to weight variation among the persons with pre-diabetes at time 2. If the estimated transition probabilities are known, these can be included so that precision bounds for domains at time 2 can be included in the constraints. An alternative is to put bounds on the maximum weight variation at time 1, and that is the approach adopted in the example in this paper. We specify that max w_{ik} / min $w_{ik} \leq W_s$ for designated sets s. For example, s may consist of persons having the same second-phase classification but different sampling weights due to different first-phase classifications.

Constraints 1, 2, and 3 guarantee that selection probabilities are between 0 and 1, and that at least one observation is taken from each non-empty sampling stratum. Some of the constraints are "box" constraints and others are linear constraints, but some of the constraints (such as 6) are nonlinear functions of $\boldsymbol{\theta}$. This necessitates use of an optimization algorithm that allows nonlinear constraints to be used.

Finally, an objective function to be minimized must be specified. The objective function will usually be related to the anticipated precision in the domains. In this paper, we use the objective function

$$f(\mathbf{n}, \mathbf{m}) = \sum_{q=1}^{Q} \alpha_q SE_q(\hat{p}),$$

where the α_q, $q = 1, \ldots, Q$ are user-specified constants specifying the relative importance of achieving the desired precision in domain q. A linear combination of variances for key quantities could also be used.

3. Software for Optimization

Many programs are available that can perform constrained nonlinear optimization, including implementations in R (R Core Team, 2016), Maple (Pintér et al., 2006), and MATLAB (Lopez, 2014). In this paper we concentrate on optimization using the OR procedure in SAS® software (SAS Institute Inc., 2015) and the Solver function in Microsoft® Excel® 2013. Each procedure has advantages and disadvantages. For

nonlinear optimization, both programs use a Newton-type method based on reducing the gradient.

Stokes and Plummer (2004) used Excel for sample design problems such as optimal allocation for stratification. Although Excel does not have the programming flexibility of some of the other languages, it has the advantage of displaying all of the features of the design so that the survey designer can immediately see what happens when constraints are tweaked. If the desired precisions cannot be met with the available budget, it is easy to change the constraints in the spreadsheet until all constraints can be met.

The generalized reduced gradient (GRG) nonlinear method in Excel must be used to optimize the nonlinear objective function. Because of the implementation in a spreadsheet, Solver computes all derivatives numerically and does not allow use of analytical derivatives; the user can specify in the options for the GRG nonlinear method whether forward or central differencing is to be used. Fylstra et al. (1998) described the algorithms used in Excel Solver.

Some users have reported that Excel Solver does not necessarily converge to a global optimum. McCullough and Heiser (2008) found that "Excel Solver has a marked tendency to stop at a point that is not a solution and declare that it has found a solution." They reported that Solver in Excel 2007 found the correct solution in 16 of 27 nonlinear regression problems tested. For some of the other problems, Solver reported that a solution had been found when in fact the gradient was nonzero. Almiron et al. (2010) found similar results for the same set of problems, with 16 problems having a correct solution with poor starting values, and 18 achieving the correct solution with better starting values. Mélard (2014) discussed issues of numerical accuracy with Excel 2010 Solver, and also interpreted the convergence statements and error messages provided by Solver.

For Excel, the Solver method must be set to "GRG Nonlinear" because this is a nonlinear optimization problem. McCullough and Wilson (1999) recommended the use of the automatic scaling option, which attempts to scale the constraints and objective function so that quantities are within limited orders of magnitude of each other, and we implemented that in all tests of the optimization. They also recommended using a convergence tolerance of 10E-7 rather than the default 10E-4. The user can also specify a maximum time limit and maximum number of iterations.

If the algorithm has found a locally optimal solution, the following message will be displayed: "Solver found a solution. All Constraints and optimality conditions are satisfied." The online help (http://www.solver.com/excel-solver-solver-found-solution-all-constraints-and-optimality-conditions-are-satisfied-0) says this means that Solver has "found a *locally optimal* solution: There is no other set of values for the decision variables *close to the current values* and satisfying the constraints that yields a better value for the objective. In general, there may be other sets of values for the variables, far away from the current values, which yield better values for the objective and still satisfy the constraints." Other possible messages are described at http://www.solver.com/excel-solver-solver-result-messages, and include "Solver could not find a feasible solution," which is interpreted as (http://www.solver.com/excel-solver-solver-could-not-find-feasible-solution-5) "this method (which always starts from the initial values of the variables) was unable to find a feasible solution; but there could be a feasible solution far

away from these initial values, which Solver might find if you run it with different initial values for the variables."

In SAS, the OPTMODEL procedure performs optimization for linear, quadratic, and general nonlinear programming problems. Because of the nonlinearity of the objective and constraint functions, the NLP (nonlinear programming) solver must be used. NLP employs an interior point algorithm (Forsgren et al., 2002; Akrotirianakis and Rustem, 2005) to optimize the objective function. If the initial values are on the boundary of the constraints, the algorithm moves them to interior points. The interior point algorithm allows some of the constraints to be violated in intermediate iterations of the algorithm. Equality constraints are incorporated as a scaled quadratic loss penalty in the objective function, where the scaling factor approaches infinity as the iterations progress.

For most objective functions and nonlinear constraints that make use of SAS library functions, SAS calculates analytic derivatives. For user-defined functions, SAS uses numerical differentiation with an option in the PROC OPTMODEL statement to use either forward or central differencing. The NLP solver rescales the objective and constraint functions dynamically, as needed.

Although SAS does not provide the same level of visibility of the results as Excel, it has several advantages for use. The constraints and starting values can be entered as data sets which makes it easy to try different values or to keep a record of the output for different input parameters. Similarly, the solutions can be output to other SAS data sets. The array notation in PROC OPTMODEL allows the same macro to be used for multiple optimizations with different values for K and Q; in Excel, changing the number of strata or domains requires the spreadsheet to be changed and all formulas rechecked. The iteration log for the SAS NLP solver provides information on constraints that are as yet unmet and the optimality error, and various diagnostics are provided when the algorithm fails to converge.

Both Excel and SAS allow for multiple starting values to be used. Using multiple starts provides some protection against getting trapped in a local minimum for the function.

4. Example

In this section, we describe a hypothetical survey and include screenshots that illustrate using Excel Solver (in Excel 2010) or PROC OPTMODEL (in SAS version 9.4) to optimize the survey design. An advantage of the Excel approach is that cells can be color coded to distinguish different components of the optimization problem. Here we use yellow to denote the inputs (e.g., misclassification probabilities, population sizes for the first-phase sampling strata, response rates, and costs), orange to denote the constraint bounds, green to denote the values that can be changed in searching for the solution (i.e., first- and second-phase sample sizes), and black to denote the objective function. Also, cell entries can be created to highlight whether or not individual constraints have been met using conditional formatting.

The survey involves a first-phase screener where one household member provides information about all other members of the household. This information is used to classify individuals into one of four first-phase sampling strata ($K = 4$), and an initial sample of persons is selected. The second-phase instrument is completed by the selected

persons themselves and these responses determine who is retained in the final sample. The survey consists of a baseline interview (and assumed follow-up interviews at later points in time).

4.1 Optimization using Excel Solver

The design challenge is to determine the optimal first- and second-phase sample sizes for the start of data collection. In this example, the optimal number of first-phase screeners is also determined at the household level. Figure 1 shows the initial assumptions about the input misclassification probabilities and population distribution across the first-phase sampling strata. In this example, the top row of the misclassification table implies that 80% of the household members classified in first-phase sampling stratum "1" will be classified as "1" at the second phase, while 10% will be classified as "2," and 5% each will be classified as "3" and "4." Note that some of the anticipated misclassification probabilities (p_{31}, p_{32}, p_{41}) are 0, and, for these, the corresponding second-phase sample sizes (m_{31}, m_{32}, m_{41}) are set equal to 0. The numbers on the right in Figure 1 provide the expected number of household members classified into each first-phase sampling stratum for every household with first-phase screening information.

Entry (i,k) of table is P(phase 2 classification is k | phase 1 sampling stratum is i)

		Phase 2 classification				Check sum	Number eligible per completed phase 1 screener
		1	2	3	4		
Phase 1 sampling stratum	1	0.8	0.1	0.05	0.05	1	0.1080
	2	0.2	0.7	0.05	0.05	1	0.2847
	3	0	0	0.9	0.1	1	0.4323
	4	0	0.05	0.1	0.85	1	1.1089

Figure 1: Misclassification probabilities and population distribution across first-phase sampling strata per completed first-phase screener.

Other cells in the Excel spreadsheet capture inputs such as the costs for the different data collection phases and constraints in terms of the overall budget, minimum final sample size, and maximum ratio of sampling weights across the 13 non-empty second-phase sampling strata. In this example, the total budget is $1,700,000, and the costs per completed first-phase screener, second-phase instrument, and baseline interview (including potential follow-up at later times) are $100, $400, and $600, respectively. The first-phase screener cost is at the household level, while the other costs are at the person level. The total final sample size must be at least 1,000 and the ratio of the maximum to minimum (nonzero) sampling weights across both phases can be at most five.

The goal of the Excel Solver computations is to find values for the number of first-phase screeners and for the first-phase and second-phase sample size vectors (**n**, **m**) that minimize the objective function and satisfy the design constraints. The inputs and results of this process are displayed in Figures 2 through 4. The green cells in Figure 2 show Solver's optimal solution for how many first-phase screeners need to be completed (2,038) and how many household members to select at the first phase from each sampling stratum (2,020 in total). The cells on the right check that the sampling probabilities do not

exceed 1. At the bottom of the figure, is another input—the assumed response rate to the second-phase screener which is 75% for persons selected at the first phase, regardless of sampling stratum.

Phase 1 sample				
Number of completed phase 1 screeners			2,038	
		Number eligible from phase 1 screener	Phase 1 sample	Sample size constraint met?
Phase 1 classification	1	220	220	TRUE
	2	580	580	TRUE
	3	881	353	TRUE
	4	2,260	866	TRUE
	Sum	3,942	2,020	
Response rate to phase 2 screener			75%	

Figure 2: Optimal first-phase sample sizes and assumed response rate to the second-phase instrument.

In this example, the optimal design selects all household members classified in sampling strata “1” and “2” at the first phase, and subsamples persons from the other two strata.

Figure 3 shows the results for the optimal number of persons to select based on the results of the second-phase instrument; or, more precisely, based on the combined results of the classifications at the first and second phases. The third column contains the misclassification probabilities (from Figure 1) presented in a list rather than table format. The “Base weight” column can be used to reflect survey weights up to the first-phase screener stage, but have been set equal to 1 here. The three columns to the right respectively reflect the sampling weights due to selection at the first phase, second phase, and both phases combined. The last column therefore captures the effects of misclassification at the first-phase screener. Not shown are cells indicating whether the second-phase sampling probabilities, overall cost, and maximum weight ratio constraints are met.)

The optimal design selects all of the second-phase respondents for the final sample from many, but not all, of the 16 second-phase sampling strata. The final column in Figure 3 shows the variation in sampling weights for persons with a given second-phase classification, depending on their first-phase sampling stratum.

Phase 1 classification	Phase 2 classification	Probability, p(ik)	Base weight	Phase 1 n responding to phase 2	Phase 2 sample	Phase 1 wt	Phase 2 wt	Sampling weight, w(ik)
1	1	0.8	1	132	132	1.0	1.0	1.0
1	2	0.1	1	17	16	1.0	1.0	1.0
1	3	0.05	1	8	4	1.0	1.9	1.9
1	4	0.05	1	8	4	1.0	2.3	2.3
2	1	0.2	1	87	86	1.0	1.0	1.0
2	2	0.7	1	305	304	1.0	1.0	1.0
2	3	0.05	1	22	11	1.0	1.9	1.9
2	4	0.05	1	22	9	1.0	2.3	2.3
3	1	0	1	-	-	2.5	0.0	0.0
3	2	0	1	-	-	2.5	0.0	0.0
3	3	0.9	1	239	239	2.5	1.0	2.5
3	4	0.1	1	27	27	2.5	1.0	2.5
4	1	0	1	-	-	2.6	0.0	0.0
4	2	0.05	1	32	32	2.6	1.0	2.6
4	3	0.1	1	65	64	2.6	1.0	2.6
4	4	0.85	1	552	552	2.6	1.0	2.6

Figure 3: Optimal second-phase sample sizes.

Figure 4 illustrates the domains of interest ($Q = 7$) and, in orange, their required minimum standard errors and sample sizes. In this example, each of the four second-phase classifications is a domain, along with classifications "1" and "2" combined, classifications "3" and "4" combined, and the entire population. The standard errors are for a population proportion of 0.5. The inputs in yellow specify the relative importance of the different domain precisions and assumed contributions to the weighting design effects from clustering of the household survey. The effective sample size column "neff" incorporates the cluster effects and the weighting effects (assuming the optimal sampling rates are used). The minimized objective function is the importance-weighted sum of standard errors for the domains of interest.

Phase 2 classification(s) forming domains	Minimum precision (SE)	Minimum sample size	Relative importance	n	Cluster deff	weff	neff	Expected SE	SE*RI
1	0.06	200	2	218	2	1.000	109	0.048	0.096
2	0.05	200	1	352	2	1.168	151	0.041	0.041
3	0.05	100	1	318	2	1.004	159	0.040	0.040
4	0.05	100	1	591	2	1.000	296	0.029	0.029
1, 2	0.04	500	2	570	2	1.119	255	0.031	0.063
3, 4	0.04	200	2	910	2	1.002	454	0.023	0.047
1, 2, 3, 4	0.02	1000	3	1,480	2	1.143	647	0.020	0.059
									0.3739

Figure 4: Precisions and sample sizes (required and estimated under optimal allocation), and objective function.

Finally, Figure 5 shows the Solver dialog window. The cell containing the objective function is referenced in the top box (along with an indication that the value is to be minimized). The green cells whose values are to be determined are entered into the next box. The orange constraints are entered into the large box; not all are visible in the screenshot. The GRG algorithm is selected due to the nonlinear nature of the optimization problem.

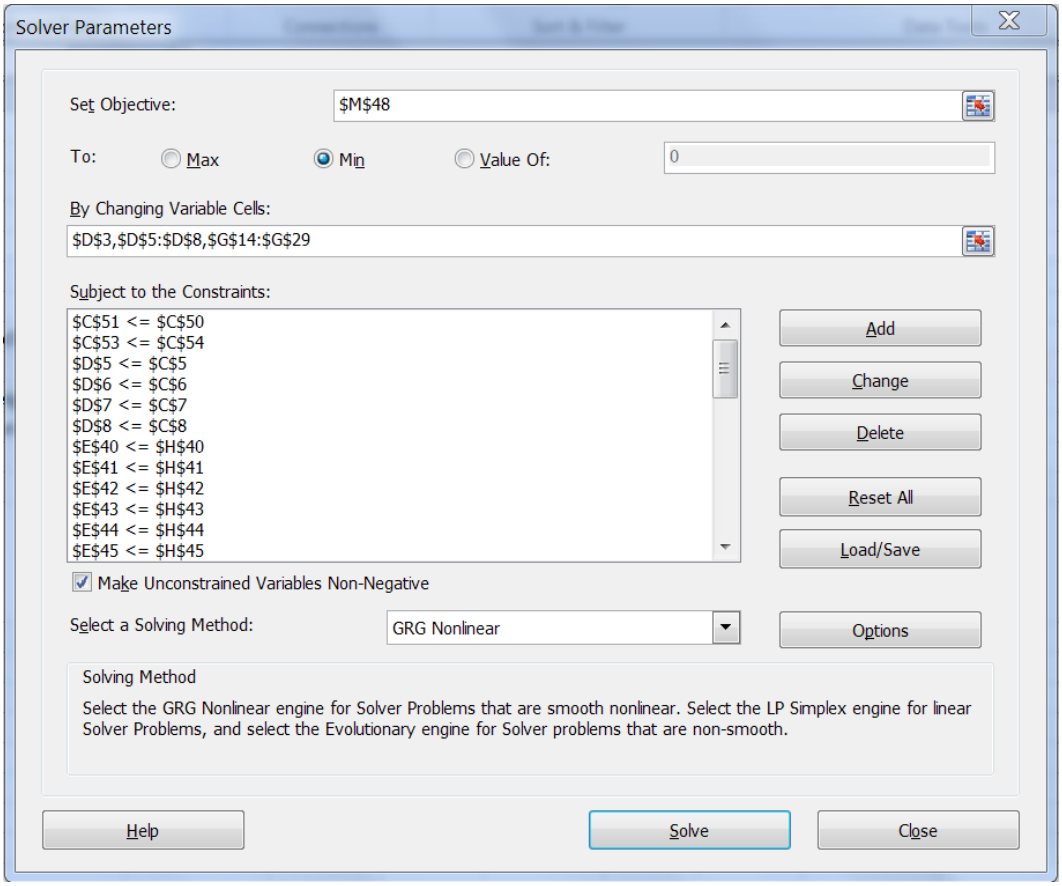

Figure 5: Example of Solver dialog window.

4.2 Optimization using SAS PROC OPTMODEL

Using SAS, the dimension of the design problem, along with several of the input parameters, can be declared using macro variables. As mentioned in Section 3, a significant advantage of this approach over Excel is the ease with which the optimization can be rescaled if, for example, the number of first-phase sampling strata needs to be increased. In Excel, this would require restructuring the spreadsheet's rows/columns and editing cell formulas, including those in the Solver specification. In SAS, the number of mutually exclusive first-phase sampling strata and the number of domains are macro arguments that can easily be changed. These parameters define the array sizes for the first- and second-phase sample sizes and misclassification probabilities. Using the current example, other inputs such as the costs, budget, assumed response rate the second-phase instrument, and maximum weight ratio can be similarly declared (or set using global parameters).

Other input values such as the misclassification probabilities, number of persons expected in each first-phase sampling stratum, mapping of sampling strata to domains, minimum desired domain precisions, and minimum desired domain sample sizes are entered using DATA steps. The READ statement can be used to load these data sets into the OPTMODEL procedure.

The sample sizes to be optimized are declared using VAR statements where starting values can be specified. Implicit variables such as the sampling weights w_{ik}, weighting effects weff_q, and standard errors $SE_q(\hat{p})$ are defined using IMPVAR statements, and the cost, precision, sample size, and weight ratio constraints are entered using CONSTRAINT statements. Finally, the MINIMIZE statement is used to request that the objective function f be minimized, and SOLVE WITH NLP tells SAS which optimization solver to use.

Figures 6 and 7 show relevant parts of the SAS listing when PROC OPTMODEL was used on the example given in this section. The results are similar, though not identical, to those obtained using Excel Solver.

Solution Summary

Solver	NLP
Algorithm	Interior Point
Objective Function	f
Solution Status	Optimal
Objective Value	0.3731016515
Optimality Error	2.9697899E-7
Infeasibility	0
Iterations	612
Presolve Time	0.00
Solution Time	2.26

Figure 6: Details of the optimization process from SAS PROC OPTMODEL.

	nhh_phase1
	2040.5

[1]	n_phase1
1	220.37
2	579.71
3	354.27
4	866.64

[1]	n_phase2
1	132.1887
2	16.5261
3	4.3520
4	3.5151
5	86.9522
6	303.9306
7	11.4741
8	9.2670
9	0.0000
10	0.0000
11	239.0849
12	25.9301
13	0.0000
14	32.4848
15	64.9847
16	552.0254

Figure 7: Optimized sample sizes using SAS PROC OPTMODEL.

5. Discussion

There are many issues to consider when designing longitudinal studies subject to misclassification. Misclassification rates affect the final yield in domains of interest and potentially lead to weight variation. The possible transition of study members between domains of interest at different times also leads to a desire to control weight variation at time 1. Both SAS PROC OPTMODEL and Excel Solver proved to be useful tools for addressing the sample design as an optimization problem. Each program has advantages that might make it more suitable for a particular survey design. PROC OPTMODEL is more flexible and is easier to use when the number of sampling strata or the number of domains may change during the survey development. The SAS program log and listing files provide valuable information on the iteration history and diagnostics for the optimization process. In practice, PROC OPTMODEL converged to a solution more often than Excel Solver when given "poor" starting values, and having two different approaches was helpful in this regard. However, being able to visualize the design assumptions, inputs, constraints, and solution in Excel Solver aids in understanding how all the design features come together, and does not require a programming background in SAS.

Many surveys are exploring the use of responsive designs in which results from the early release groups are used to modify the design for the later release groups. This is often done in order to improve survey cost efficiency and reduce bias (Groves and Heeringa, 2006). Similarly, the programs used in this paper can be re-run with updated estimates for inputs such as the misclassification probabilities and response rates based on survey information that has accrued. The sample design is adapted in an effort to meet desired precision objectives for persons in domains of interest and other survey goals.

Acknowledgements

Microsoft® and Excel® are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. This article is an independent publication and is not affiliated with, nor has it been authorized, sponsored, or otherwise approved by Microsoft Corporation. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

References

Akrotirianakis, I., and Rustem, B. (2005). Globally convergent interior-point algorithm for nonlinear programming. *Journal of Optimization Theory and Applications*, 125(3), 497-521.

Almiron, M.G., Lopes, B., Oliveira, A.L.C., Medeiros, A.C., and Frery, A.C. (2010). On the numerical accuracy of spreadsheets. *Journal of Statistical Software*, 34, 1-29.

Boonstra, H. (2001). The 'Add Health' survey: Origins, purposes, and design. *The Guttmacher Report on Public Policy*, 4(3), 10-12.

Carolina Population Center (2015). The National Longitudinal Study of Adolescent to Adult Health: Wave V. http://www.cpc.unc.edu/projects/addhealth/design/wave-v-1, last accessed December 19, 2015.

Forsgren, A., Gill, P.E., and Wright, M.H. (2002). Interior methods for nonlinear optimization. *SIAM Review*, 44(4), 525-597.

Fylstra, D., Lasdon, L., Watson, J., and Waren, A. (1998). Design and use of the Microsoft Excel Solver. *Interfaces*, 28(5), 29-55.

Groves, R.M., and Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169(3), 439-457.

Harris, K.M. (2007). Design features of Add Health. *Social Psychology Quarterly*, http://www.asanet.org/journals/spq/health.cfm, last accessed December 19, 2015.

Harris, K.M., Halpern, C.T., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., and Udry, J.R. (2015). The National Longitudinal Study of Adolescent Health: Research Design. http://www.cpc.unc.edu/projects/addhealth/design, last accessed December 19, 2015.

Kalton, G., Brick, J.M., and Lê, T. (2005). Estimating components of design effects for use in sample design. In *Household Sample Surveys in Developing and Transition Countries*, New York: United Nations, available at http://unstats.un.org/unsd/hhsurveys/pdf/Chapter_6.pdf, last accessed August 1, 2014.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8(2), 183-200.

Lopez, C. (2014). *MATLAB optimization techniques*. New York: Apress.

McCullough, B.D. and Heiser, D.A (2008). On the accuracy of statistical procedures in Microsoft Excel 2007. *Computational Statistics & Data Analysis*, 52(10), 4570-4578.

McCullough, B.D., and Wilson, B. (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics & Data Analysis*, 31(1), 27-37.

Mélard, G. (2014). On the accuracy of statistical procedures in Microsoft Excel 2010. *Computational Statistics*, 29(5), 1095-1128.

Morton, K.B., Martin, P.C., Shook-Sa, B.E., Chromy, J.R., and Hirsch, E.L. (2013). 2012 National Survey on Drug Use and Health: Sample Design Report. Research Triangle Park, NC: RTI International, http://www.samhsa.gov/data/sites/default/files/NSDUH2012MRB-Ammended/NSDUHmrbSampleDesign2012.pdf, last accessed December 19, 2015.

Pintér, J.D., Linder, D., and Chin, P. (2006). Global Optimization Toolbox for Maple: An introduction with illustrative applications. *Optimisation Methods and Software*, 21(4), 565-582.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, www.r-project.org.

SAS Institute Inc. (2015). *SAS/OR® 14.1 User's Guide: Mathematical Programming*. Cary, NC: SAS Institute Inc.

Stokes, L., and Plummer, J. (2004). Using spreadsheet solvers in sample design. *Computational Statistics & Data Analysis*, 44(3), 527-546.