Optimal Sampling Fractions for Two-Phase Sampling for Nonresponse in the Real World

Barbara Lepidus Carlson
Mathematica Policy Research
955 Massachusetts Avenue, Suite 801, Cambridge, MA 02139

Abstract

Two-phase sampling is a long-standing sampling method. It identifies subpopulations of interest in the first phase of a survey, from which a random subsample is selected in the second phase for further data collection, using the new information to further stratify or to narrow the survey population to a particular subgroup. It is also used to randomly subsample survey nonrespondents for more intensive follow-up. In this context, the phase-one nonrespondents are considered a subpopulation that is identified after data collection efforts have been completed with the initial mode and protocol. The more intensive phase-two data collection protocol is generally more expensive to implement than the first and is expected to have a greater success rate. However, budgetary constraints generally limit how many nonrespondents data collectors can attempt to contact using this more expensive protocol. Hansen et al.'s (1953) work provides optimal values for the fraction of phase-one nonrespondents to be subsampled for phase two ($1/k$) and for the initial sample size (M) in a two-phase sample with a subsample of proportion $1/k$. However, these calculations assume that phase-two methods result in 100 percent response, which is not often the case in real-world scenarios. In this paper, I derive new optimal values for M and k under the more realistic scenario in which not all phase-two attempts result in a response.

Key Words: two-phase sampling, double sampling, nonresponse

1. Introduction

Two-phase sampling is an efficient and cost-effective method to increase survey response rates. In this method, initial survey nonrespondents are randomly subsampled for more intensive follow-up in a second phase of data collection that uses a different and generally more expensive protocol.

Consider two interviewer-administered data collection protocols: (1) a telephone interview and (2) an in-person interview. A telephone interview is less expensive per completed interview than an in-person interview, but the latter is generally more effective at gaining respondents' cooperation and obtaining completed interviews.[1] In this scenario, in-person interviewing might not be an option for all sample members, given cost constraints. In a two-phase sample design, interviewers would attempt to contact all sample members initially using the phase-one protocol (telephone interviewing). After a fixed and well-defined level of effort adhering to the initial data collection protocol, researchers would determine who responded and who did not, then randomly subsample phase-one nonrespondents and attempt to complete the survey in person for this subsample only (phase two). At this point, all efforts to contact any phase-one nonrespondents not

[1] Other examples might include a web survey in phase one and telephone survey in phase two, or a lower incentive amount in phase one and higher amount in phase two.

subsampled for phase two would cease. This method is clearly less expensive than attempting to contact all phase-one nonrespondents using in-person interviewing.

Two-phase sampling is not to be confused with *two-stage* sampling, the latter referring to a clustered or nested sample design. However, two-phase sampling can be used in conjunction with various simple and complex sampling methods. For illustrative purposes, this paper will use a simple random sample to demonstrate the methodology.

2. Background

Two-phase sampling (also known as *double sampling*) is a concept with a long history. Neyman (1938) described it as a way to sample a specific subpopulation or stratify on a certain characteristic when the variables that define the desired subpopulation or strata are not available on the initial sample frame. In this scenario, researchers collect data from a sample survey to obtain the information needed for a more targeted sample design in phase two. The new information helps researchers refine the sample—either to narrow the population to a particular subgroup or to stratify and perhaps oversample certain subgroups. Data collectors then attempt to interview a random subsample of this targeted population.

In 1946, Hansen and Hurwitz adapted Neyman's methodology to address the problem of survey nonresponse. But instead of using phase one to obtain characteristics not available on the initial sample frame, they proposed using phase-one data collection to identify which sample members *do not respond to this kind of survey using the phase-one protocol*, essentially considering response as an intrinsic characteristic of the sample member. Then, in phase two, researchers select a subsample of these phase-one nonrespondents and attempt to contact them using a different protocol, a more expensive method with an expected higher success rate overall and, in particular, which converts phase-one nonrespondents to respondents.

3. Response Rates

As mentioned previously, the main goal of two-phase sampling for nonresponse is to increase the response rate in a cost-effective manner. For simplicity, I use a simple sample design (single stage, no stratification) in which we know all sample members are eligible. In this situation, the response rate is calculated to be

$$RR = \frac{R}{M}$$

where $M = R + N$ is the total sample size, R is the number of respondents, and N is the number of nonrespondents.

If we implement a two-phase design in which the subscript 1 represents the results after phase one, then the phase-one response rate is

$$RR_1 = \frac{R1}{M}$$

where $M = R1 + N1$ is the total sample size, $R1$ is the number of phase-one respondents, and $N1$ is the number of phase-one nonrespondents.

If we randomly subsample $1/k$ of the $N1$ phase-one nonrespondents for phase two, then the cumulative response rate through phase two is calculated to be either of the following two algebraically equivalent formulas:

$$RR = \frac{R1 + (k)R2}{R1 + (k)(R2 + N2)} = RR_1 + (1 - RR_1)\frac{R2}{R2 + N2}$$

where $R2$ is the number of respondents to phase two among the subsampled phase-one nonrespondents, $N2$ is the number of nonrespondents to phase two among the subsampled phase-one nonrespondents, and $R2 + N2 = (1/k)(N1)$.

The subsampling weight is represented by k, the inverse of the subsampling fraction. It indicates the number of phase-one nonrespondents in the sample that each subsampled case represents. In two-phase sampling, the subsampling weight k must be used to make estimates or calculate response rates to properly account for the entire sample.

4. Example

To demonstrate the procedure, I introduce a simple example that I will use throughout the paper. In this example, the initial sample is M = 2,000, and $R1$ = 500 respond to phase one—a telephone interview with 12 attempts made over four weeks. The phase one response rate is as follows:

$$RR_1 = \frac{500}{500 + 1500} = .25$$

Next, suppose we randomly subsample one-third of the $N1$ phase-one nonrespondents. If $1/k = .33$, then we subsample 500 of the 1,500 phase-one nonrespondents for phase two (in this example, in-person visits). If 250 of these 500 are converted to respondents in phase two and 250 are persistent nonrespondents, the phase-two response rate is 50 percent, and the cumulative response rate is as follows:

$$RR = \frac{500 + (3)250}{500 + (3)(250 + 250)} = .25 + (1 - .25)\frac{250}{250 + 250} = .625$$

This is quite an increase in the response rate. Had we stopped data collection after phase one, the response rate would have been only 25 percent. By progressing to phase two, the response rate jumped to 62.5 percent. As I will demonstrate later in this paper, this jump in response rate would be the same regardless of the proportion subsampled for phase two. It is important to note that the 1,000 phase-one nonrespondents who were *not* subsampled are no longer included in subsequent calculations, as they are now represented by those who *were* subsampled when properly weighted by k.

5. Cost

It can be expensive to obtain this increase in response rate in the phase-two data collection, but costs can be substantially reduced by subsampling phase-one nonrespondents for phase two, rather than sending all phase-one nonrespondents to phase two. Suppose C_1 is the cost per completed interview using the phase-one protocol, and C_2 is the cost per completed interview using the phase-two protocol, with C_2 greater than C_1. (To keep things simple, we assume here that C_1 does not vary for one-phase and two-phase designs.) Continuing with the example—and assuming the same two response rates for phases one and two—suppose C_1 = \$100 and C_2 = \$500. Table 1 shows the response rate and the cost for three designs: (1) a phase-one-only design, (2) a two-phase design in which all phase-one nonrespondents go to phase two, and (3) a two-phase design in which a subsample of one-third of nonrespondents go to phase two.

Table 1: Comparison of Cost and Response Rate Across Three Designs

Design	*M*	*R*1	*R*2	*N*	*RR*	Cost (in thousands)
Phase one only	2,000	500	0	500	.250	$50
Phase one + all nonrespondents go to phase two	2,000	500	750	1,250	.625	$425
Phase one + one-third of nonrespondents go to phase two	2,000	500	250	750	.625	$175

This table shows that both two-phase designs have the same response rate of 62.5 percent, but subsampling in phase two reduces the cost by $250,000. This third design is more costly than the phase-one-only design, but results in a higher number of completed interviews and much higher response rate.

6. Sample Size and Precision

There is also a price to pay to conduct phase-two subsampling, in terms of precision. As the previous example demonstrates, the total completed sample size is substantially lower than it would be had all phase-one nonrespondents moved to phase two (750 versus 1,250). In addition to the smaller sample size, there is a weighting design effect because each respondent to the phase-two data collection has a weight that is *k* times that of a phase-one respondent. This variability in the weights adversely impacts the precision of estimates. As I will discuss later in the paper, increasing the response rate is an attempt to reduce nonresponse bias in our estimates. But in doing so, we increase the variance and thus encounter the bias-variance trade-off that statisticians often face.

A design effect is a measure of the impact of a complex sample design on the variance of estimates. Complexities can include unequal weighting (due to unequal sampling probabilities or weighting adjustments), clustering, and stratification. Unequal weighting and clustering tend to increase the variance of estimates relative to a simple random sample. A design effect is the ratio of the true variance (properly accounting for design complexities) to the variance one would obtain for a simple random sample of the same nominal sample size. A design effect of 1.5 indicates that a design increased the variance of an estimate by 50 percent and effectively reduced the sample size by one-third.

Because the subsampling weight of *k* is applied to some sample members (those who were phase-one nonrespondents but then responded in phase two) and not others (phase-one respondents), weighting disparities are introduced, which in turn introduce a design effect (or an additional design effect if the original design was already complex). I will introduce specific calculations later in the paper, but in Table 2 I add a column showing the effective sample size in each of the three designs, where the effective sample size is the nominal sample size divided by the newly introduced design effect. The first two designs have a design effect of 1, as no subsampling occurred for phase two.

Table 2: Comparison of Precision Across Three Designs

Design	*M*	*N*	**Effective sample size**	*RR*	Cost (in thousands)	Cost per effective sample unit	Relative precision[b] (percent)
Phase one only	2,000	500	**500**	.250	$50	$100	---
Phase one + all nonrespondents go to phase two	2,000	1,250	**1,250**	.625	$425	$340	36.8
Phase one + one-third of nonrespondents go to phase two	2,000	750	**568**[a]	.625	$175	$308	6.2

[a] Design effect = 1.32.
[b] *Relative precision* is defined here as the reduction in the size of a 95 percent confidence interval (half width) for a proportional outcome of 0.5, relative to that of the phase-one only design.

The third design (subsampling for phase two) has an effective sample size less than half that of the second design. Combining the impacts of the various designs on cost and effective sample size, the cost per effective sample unit is slightly lower for the one-third subsample in phase two, relative to that of the design in which all nonrespondents go to phase two ($308 versus $340). In terms of precision, the subsample design does not improve precision very much over the phase-one-only design, with the latter having a 95 percent confidence interval (around a proportion of 0.5) of plus or minus .044, and the former having an interval of plus or minus .041, due to the slightly larger effective sample size.

7. Moving Parts

Thus, there are three moving parts to compare: (1) response rate, (2) cost, and (3) sample size and precision. To discuss these moving parts, I return to the concept of response rates. Historically, response rates were used as a barometer for the risk of nonresponse bias. The higher the response rate, the lower the risk for nonresponse bias. But in recent years, this assumption has received more thought. Although nonresponse bias rarely can be measured directly, in theory, one can represent nonresponse bias by

$$\text{Bias} = (\bar{y}_r - \bar{y}_{nr}) \cdot \frac{nr}{n}$$

The first term represents the difference in the mean value of outcome *y* between the respondents and the nonrespondents, and the second term represents the nonresponse rate (one minus the response rate). As the nonresponse rate approaches zero (as the response rate approaches one), the bias approaches zero. However, as the difference in the outcome between respondents and nonrespondents approaches zero, the bias also approaches zero, regardless of the nonresponse rate. If the propensity to respond is unrelated to key outcomes, then a low response rate is less problematic in terms of the risk for nonresponse bias. But many data collection clients (such as government agencies and foundations) and professional journals consider the response rate an important measure of the quality of survey estimates. That is the underlying premise when using two-phase sampling to

increase the response rate as cost-effectively as possible. In fact, increasing the response rate this way, through random subsampling, is more likely to reduce nonresponse bias than by struggling to get responses from the next easiest cases among the entire sample.

The precision of estimates (standard errors and confidence intervals) and, related to this, the power to detect differences, are affected by the total number of responses and the design effect; that is, the effective sample size. One can increase the effective sample size by increasing the nominal sample size or decreasing the design effect or both. The choice of the subsampling fraction has an impact on both of these, as well as on the cost.

When comparing the two-phase subsampling approach to a design with only one phase of data collection, each with the same initial sample size, the two-phase subsampling approach yields more completed interviews and at a higher response rate, but also at a higher cost. For the same initial sample size, the effective sample size could be larger or smaller than the phase-one sample size, depending on the subsampling fraction and the response rates in each of the two phases.

When comparing the two-phase subsampling approach to a design in which all phase-one nonrespondents go to phase two, the subsampling approach yields fewer completed interviews and has an even smaller effective sample size, but has a lower cost and yields the same response rate. This is true regardless of the subsampling fraction and the response rates in each of the two phases. To achieve the same effective sample size but still with a lower cost, one can increase the initial sample size for two-phase subsampling. Similarly, one can increase the initial sample size for two-phase subsampling to increase the effective sample size at the same cost as the design in which all phase one nonrespondents go to phase two.

8. Subsampling Fraction

As I have shown, after fixing the two protocols for phases one and two, along with their expected response rates, the overall response rate is fixed for a two-phase design, regardless of the subsampling fraction. However, if we fix cost, the effective sample size will vary depending on the subsampling fraction and the initial sample size. Similarly, if we fix the effective sample size, the cost will vary depending on the subsampling fraction and the initial sample size.

According to Hansen et al. (1953), the optimal value for k (where $1/k$ is the subsampling fraction for phase two) is

$$k = \sqrt{\frac{C_2 RR_1}{C_0 + C_1 RR_1}}$$

where C_0 is the cost per attempt in phase one, C_1 is the cost per complete in phase one, and C_2 is the cost per complete in phase two.

Optimal here means that this value of k is best if one fixes cost and maximizes the effective sample size, or if one fixes the effective sample size and minimizes cost. If $C_0 = 0$, or if C_0 is included in C_1, then k simplifies to

$$k = \sqrt{\frac{C_2}{C_1}}$$

9. Deriving Existing Formulas Assuming Full Response in Phase Two

According to Hansen et al. (1953), the assumption is that all those subsampled for phase two will result in completes. Although perhaps that was a fair assumption in 1953, it is no longer reasonable to assume 100 percent response among subsampled phase-one nonrespondents when attempting to contact using the phase-two protocol. In this paper, I expand on the original formulas to accommodate nonresponse in phase two. To do so, I first replicate their formulas for cost, initial sample size, number of completes, design effects, and the optimal value of *k*, making their assumption of complete response in phase two.

If the initial sample size is *M* and we subsample 1/*k* for phase two, the number of completes (*N*) in a two-phase design with full participation in phase two is:

$$N = R1 + R2 = M(RR_1 + (1 - RR_1)/k)$$

and the associated cost is:

$$cost = M(C_0 + C_1 RR_1 + C_2(1 - RR_1)/k)$$

The design effect due to unequal weighting (*deff*) is:

$$deff = \frac{N\sum WT^2}{(\sum WT)^2} = \frac{M(RR_1 + (1 - RR_1)/k)(M \cdot RR_1 \cdot 1^2 + M \cdot (1 - RR_1)/k \cdot k^2)}{(M \cdot RR_1 \cdot 1 + M \cdot (1 - RR_1)/k \cdot k)^2}$$

$$= \frac{(RR_1 + (1 - RR_1)/k)(RR_1 + (1 - RR_1) \cdot k)}{(RR_1 + (1 - RR_1))^2}$$

$$= (RR_1 + (1 - RR_1)/k)(RR_1 + (1 - RR_1) \cdot k)$$

The effective sample size (*effn*), which is the nominal sample size divided by the design effect, is:

$$effn = \frac{M(RR_1 + (1 - RR_1)/k)}{(RR_1 + (1 - RR_1)/k)(RR_1 + (1 - RR_1) \cdot k)} = \frac{M}{(RR_1 + (1 - RR_1) \cdot k)}$$

9.1 Matching Effective Sample Size and Minimizing Cost

When all phase-one nonrespondents go to phase two (and all respond), the effective sample size is equal to the nominal sample size selected. Suppose the initial sample size (and the number of respondents) for such a design is M_0. If we want to match the effective sample size for phase two subsampling to the effective sample size for the two-phase design with no subsampling, then we would have to inflate M_0 by a certain factor. To derive that factor, set the two effective sample sizes equal to each other:

$$M_0 = \frac{M}{(RR_1 + (1 - RR_1) \cdot k)}$$

then $$M = M_0(RR_1 + (1 - RR_1) \cdot k) = M_0\big(1 + (1 - RR_1)(k - 1)\big)$$

The last term above is what Hansen et al. indicate, but I will use the preceding term for the calculations that follow. The inflation factor is then $(RR_1 + (1 - RR_1) \cdot k)$. To find the

optimal value of k, we minimize the cost for this fixed effective sample size when subsampling for phase two:

$cost = M(C_0 + C_1RR_1 + C_2(1 - RR_1)/k) = \quad M_0(RR_1 + (1 - RR_1) \cdot k)(C_0 + C_1RR_1 + C_2(1 - RR_1)/k)$

To minimize the cost function, we factor out constant M_0, take the derivative of this equation with respect to k,[2] set equal to zero, and solve for k:

$$k = \sqrt{\frac{C_2RR_1}{C_0 + C_1RR_1}}$$

9.2 Matching Cost and Maximizing Effective Sample Size

Similarly, if we set the cost for the two-phase design with subsampling to that of the two-phase design with no subsampling, then

$$M_0\big(C_0 + C_1RR_1 + C_2(1 - RR_1)\big) = M\left(C_0 + C_1RR_1 + \frac{C_2(1-RR_1)}{k}\right)$$

and then $$M = M_0 \frac{\big(C_0+C_1RR_1+C_2(1-RR_1)\big)}{\left(C_0+C_1RR_1+\frac{C_2(1-RR_1)}{k}\right)}$$

Next, we try to maximize the effective sample size for this value of M, which is the same as minimizing the inverse of the effective sample size:

$$\frac{M_0}{effn} = \frac{(C_0 + C_1RR_1 + C_2(1 - RR_1)/k)(RR_1 + (1 - RR_1)k)}{\big(C_0 + C_1RR_1 + C_2(1 - RR_1)\big)}$$

If we again factor out constant M_0, take the derivative with respect to k,[3] set to zero, and solve for k, we once again find that the optimal value of k is

$$k = \sqrt{\frac{C_2RR_1}{C_0 + C_1RR_1}}$$

10. Deriving New Formulas Allowing for Nonresponse in Phase Two

I now introduce a new term, P_2, which is equal to $(1\text{-}RR_1)(RR_2)$, where RR_2 is the response rate in phase two. Using this, we can derive new formulas allowing for less than full response in phase two. If we again assume the initial sample size is M and we subsample $1/k$ for phase two, the number of completes in a two-phase design with some nonresponse in phase two is

$completes = N = M(RR_1 + (1 - RR_1)RR_2/k) = M(RR_1 + P_2/k)$

[2] $\frac{d(cost)}{dk} = \big((1 - RR_1) \cdot (C_0 + C_1RR_1)\big) - \big(C_2 \cdot RR_1 \cdot (1 - RR_1)\big)k^{-2}$

[3] $\frac{d\left(\frac{M_0}{effn}\right)}{dk} = \left(\frac{(C_0+C_1RR_1)(1-RR_1)}{(C_0+C_1RR_1+C_2(1-RR_1))}\right) - \left(\frac{C_2RR_1(1-RR_1)}{(C_0+C_1RR_1+C_2(1-RR_1))}\right)k^{-2}$

and the associated cost is

$$cost = M(C_0 + C_1 RR_1 + C_2(1 - RR_1)RR_2/k) = M(C_0 + C_1 RR_1 + C_2 P_2/k)$$

The design effect due to unequal weighting is

$$deff = \frac{N\sum WT^2}{(\sum WT)^2} = \frac{M(RR_1 + P_2/k)(M \cdot RR_1 \cdot 1^2 + M \cdot P_2/k \cdot k^2)}{(M \cdot RR_1 \cdot 1 + M \cdot P_2/k \cdot k)^2}$$

$$= \frac{(RR_1 + P_2/k)(RR_1 + P_2 \cdot k)}{(RR_1 + P_2)^2}$$

The effective sample size, which is the nominal sample size divided by the design effect, is

$$effn = \frac{M(RR_1 + P_2/k)(RR_1 + P_2)^2}{(RR_1 + P_2/k)(RR_1 + P_2 \cdot k)} = \frac{M(RR_1 + P_2)^2}{(RR_1 + P_2 \cdot k)}$$

When all phase-one nonrespondents go to phase two, the effective sample size is equal to the nominal sample size times the cumulative response rate; that is, there is no subsampling design effect.

10.1 Matching Effective Sample Size and Minimizing Cost

If we want to match the effective sample size for phase-two subsampling to the effective sample size for the two-phase design with no subsampling, then we have to inflate initial sample size M_0 by a certain factor. To derive that factor,

set: $M_0(RR_1 + P_2) = \frac{M(RR_1+P_2)^2}{(RR_1+P_2 \cdot k)}$

then $M = M_0 \frac{(RR_1+P_2 \cdot k)}{(RR_1+P_2)}$

The inflation factor is then $(RR_1 + P_2 \cdot k)/(RR_1 + P_2)$. To find the optimal value of k, we minimize the cost for a fixed effective sample size:

$$cost = M(C_0 + C_1 RR_1 + C_2 P_2/k) = M_0 \frac{(RR_1+P_2 \cdot k)}{(RR_1+P_2)}(C_0 + C_1 RR_1 + C_2 P_2/k)$$

If we factor out constant M_0, take the derivative of this equation with respect to k,[4] set to zero, and solve for k, we again get the same formula:

$$k = \sqrt{\frac{C_2 RR_1}{C_0 + C_1 RR_1}}$$

[4] $\frac{d(cost)}{dk} = \left(\frac{C_0 P_2 + C_1 P_2 RR_1}{RR_1+P_2}\right) - \left(\frac{C_2 P_2 RR_1}{RR_1+P_2}\right) k^{-2}$

10.2 Matching Cost and Maximizing Effective Sample Size

Similarly, if we set the cost for the two-phase design with subsampling to that for the two-phase design with no subsampling, then maximize the effective sample size, then:

$$M_0(C_0 + C_1RR_1 + C_2P_2) = M\left(C_0 + C_1RR_1 + \frac{C_2P_2}{k}\right)$$

and then $$M = M_0 \frac{(C_0+C_1RR_1+C_2P_2)}{\left(C_0+C_1RR_1+\frac{C_2P_2}{k}\right)}$$

Next, we try to maximize the effective sample size for this value of M, which is the same as minimizing the inverse of the effective sample size (once again, factoring out constant M_0):

$$\frac{M_0}{effn} = \frac{(C_0 + C_1RR_1 + C_2P_2/k)(RR_1 + P_2k)}{(C_0 + C_1RR_1 + C_2P_2)(RR_1 + P_2)^2}$$

If we take the derivative of this formula with respect to k,[5] set to zero, and again solve for k, the optimal value for k is the same:

$$k = \sqrt{\frac{C_2RR_1}{C_0 + C_1RR_1}}$$

10.3 Matching Other Designs

I went through similar steps, finding the optimal value of k, matching the initial sample size for the two-phase subsample design to either the effective sample size or the cost for a phase-one only design. Again, although the initial sample size inflation factors changed, the optimal value of k remained the same. This led me to try to find a generalized formula to prove that this value of k is always optimal. To do this, we have to assume that the initial sample size for the two-phase subsample design is a function of k or $1/k$. Otherwise, with no other constraints, the optimal value of $1/k$ to minimize cost is 0. However, if the initial sample size can be written in the following format

$$M = M_0 \cdot F \cdot (RR_1 + P_2k)$$

where F is any factor unrelated to k, then the derivative of the associated cost function would be

$$\frac{d(cost)}{dk} = P_2F(C_0 + C_1RR_1) - P_2FC_2RR_1k^{-2}$$

If we set this equal to zero and solve for k, the optimal value of k is the same. Doing the same exercise but trying to maximize the effective sample size proves even less generalizable. To solve the formula, we would have to use the following equation format for the initial sample size:

[5] $\frac{d\left(\frac{M_0}{effn}\right)}{dk} = \left(\frac{(C_0+C_1RR_1)P_2}{(C_0+C_1RR_1+C_2P_2)(RR_1+P_2)^2}\right) - \left(\frac{C_2RR_1P_2}{(C_0+C_1RR_1+C_2P_2)(RR_1+P_2)^2}\right)k^{-2}$

$$M = M_0 \left(\frac{C_0 + C_1 RR_1 + G}{C_0 + C_1 RR_1 + C_2 P_2 / k} \right)$$

where *G* is any factor unrelated to *k*.

11. Example (Continued)

Returning to the example, assuming that $C_0 = 0$, the optimal value of *k* would be $\sqrt{C_2/C_1}$=$\sqrt{500/100}$=2.236, which points to an optimal sampling fraction of .4472. If we apply the sample size inflation factor for minimizing cost relative to the two-phase design without subsampling with an initial sample size of 2,000, the initial sample size is 3,483 for the two-phase subsampling design. If we apply the sample size inflation factor for maximizing the effective sample size, keeping the cost constant, the initial sample size is 3,904 (Table 3).

Table 3: Comparison of Precision When Using Optimal *k*

Design	*M*	*N*	Effective sample size	*RR*	Cost (in thousands)	Cost per effective sample unit	Relative precision[a] (percent)
Phase one only	2,000	500	500	.250	$50	$100	---
Phase one + all nonrespondents go to phase two	2,000	1,250	**1,250**	.625	**$425**	$340	36.8
Phase one + .4472 of nonrespondents go to phase two (design effect of 1.164)	2,000	835	718	.625	$218	$303	16.6
	3,483	1455	**1,250**	.625	$379	$303	36.8
	3,904	1631	1,401	.625	**$425**	$303	40.3

[a] *Relative precision* is defined here as the reduction in the size of a 95 percent confidence interval (half width) for a proportional outcome of 0.5, relative to that of the phase-one-only design.

If we initially sample 3,483 and subsample .4472 of phase-one nonrespondents, the response rate (.625) and effective sample size (1,250) are the same as when all phase-one nonrespondents (among an initial sample size of 2,000) go to phase two, but for a lower cost: $379,000 compared to $425,000. Similarly, if we initially sample 3,904 and subsample .4472 of phase-one nonrespondents, the response rate (.625) and cost ($425,000) are the same as those in the comparison design, but with a higher effective sample size: 1,401 compared to 1,250. For all designs with the optimal subsampling fraction, the cost per effective sample unit is $303, which is lower than the $340 per unit for the design in which all phase-one nonrespondents go to phase two.

12. Discussion

In addition to developing new formulas related to two-phase subsampling for nonresponse, this paper shows that the formula for the optimal subsampling fraction appears to be constant across various types of constraints. However, there are some key points to emphasize when using two-phase sampling for nonrespondents. To begin, this method is most useful when one of the key goals is to maximize the response rate while controlling costs. It might not be suitable for all data collections, depending on schedule and possible methods and modes. It is used for the American Community Survey (U.S. Census Bureau 2014), where phase two involves in-person interviewing. The American Association for Public Opinion Research (2015) includes a section on how to calculate response rates for two-phase designs in its *Standard Definitions* document. When using this method, it is absolutely essential to incorporate the subsampling weight k in response rates, weights, and estimates.

Subsampling rules and mechanisms should ideally be built into the sample design from the beginning but can also be implemented after data collection has started. Various mechanisms are available for the subsampling, the most obvious being the use of random numbers or a random selection of the original sample or of the phase-one nonrespondents. If the sample has been divided into random replicates for regulating sample releases over time, these random replicates can serve a second purpose, determining which phase-one nonrespondents go to phase two.

After the fixed level of effort under the phase-one methodology, all efforts to contact the nonrespondents who were not subsampled for phase two must cease, and any completed interviews from them that might materialize later (mail-in surveys, web surveys, and incoming calls) cannot be used. The fixed level of effort under phase one should be selected carefully. Staying in phase one too long could waste money, but leaving phase one too early could unnecessarily reduce the effective sample size. The time dimension of the level of effort cutpoint can be relative to each case (on a rolling basis) rather than fixed to a specific calendar date, when the first attempt for all cases is not on the same date.

Those who use the formulas in this paper must specify the cost per attempt (if applicable), the cost per complete using the phase-one method, and the cost per complete using the phase-two method. It is important to note that as C_2 approaches C_1, the optimal value of k decreases and the subsampling fraction increases. Similarly, as C_2 increases relative to C_1, the optimal subsampling fraction decreases.

Further, users of these formulas must specify the phase-one response rate and the phase-two response rate. The overall response rate for a two-phase sample will be the same with any subsampling fraction. Technically, the phase-two response rate is a conversion response rate; that is, among those who data collectors attempted to contact in phase one but who did not respond, what is the response rate using the phase-two methodology? For the two-phase methodology to successfully increase the overall response rate, the phase-two response rate can be lower than the phase-one response rate, as long as it is greater than zero. The response rate formulas in this paper do not fully account for issues of eligibility (including undetermined eligibility status) or harsh refusals or for other sample members whom data collectors might not attempt to contact in phase two. Although the reported response rates should incorporate all of these intricacies, the main purpose of calculating the response rates here are to have "apples-to-apples" comparisons between different designs, in which case such factors are probably not important to incorporate and doing so would substantially increase their complexity.

13. Next Steps

As mentioned earlier in this paper, phase-two subsampling for nonresponse can be used alongside more complex sample designs, such as stratification (with or without oversampling), probability proportional to size sampling, geographic or other clustering, and other designs. Design effects due to unequal weighting in the original sample design would need to be incorporated into the formulas for effective sample sizes. These same complex design techniques can be used when subsampling for phase two, including using response propensity scores or *R*-indicators ("representativeness" indicators) (Schouten et al. 2012) to better target the nonrespondents for phase-two subsampling. The formulas in this paper can be tailored to individual design complexities.

Other enhancements to the formulas can include allowing C_1 to be different for phase-one-only designs (where cases might be worked longer and more intensively); adding fixed costs (such as additional statistical labor for sampling and weighting); factoring in additional sample for two-stage clustered designs and the impact on the design effect due to clustering; and incorporating additional design effects due to weighting, with nonresponse weighting adjustments possibly varying more when response rates are lower. Study designers can include all of these as optional enhancements to the formulas presented in this paper—in addition to using formulas for confidence intervals and minimum detectable differences—to see the impact of various two-phase designs on precision in practical terms.

Acknowledgements

The author acknowledges her colleague Frank Potter, who provided helpful comments for this paper.

References

American Association for Public Opinion Research. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. *8th edition.* Deerfield, IL: American Association for Public Opinion Research, 2015.

Hansen, Morris H., and William N. Hurwitz. "The Problem of Non-Response in Sample Surveys." *Journal of the American Statistical Association,* vol. 41, no. 236, 1946, pp. 517–529.

Hansen, Morris H., William N. Hurwitz, and William G. Madow. *Sampling Survey Methods and Theory, Volume I, Methods and Applications.* New York: Wiley, 1953.

Neyman, J. "Contributions to the Theory of Sampling Human Populations." *Journal of the American Statistical Association,* vol. 33, no. 201, 1938, pp. 101–116.

Schouten, B., J. Bethlehem, K. Buellens, O. Kleven, G. Loosveldt, A. Luiten, K. Rutar, N. Shlomo, and C. Skinner. "Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-indicators and Partial R-indicators." *International Statistical Review,* vol. 80, no. 3, 2012, pp. 382–399.

U.S. Census Bureau. "American Community Survey Design and Methodology (Version 2.0)." Washington, DC: U.S. Census Bureau, January 2014.