Extensions to the Two-stratum Model for Sampling Rare Subpopulations in Telephone Surveys

Walter R. Boyle[1] and William D. Kalsbeek[2]
RTI International[1]
University of North Carolina-Chapel Hill, Dept of Biostatistics[2]

Abstract

Population subgroups defined by demographic and other characteristics are often an important focal point of samples in telephone surveys. We consider a class of two-stratum telephone sample designs where part of the frame with higher subgroup concentration (the first stratum) is disproportionately sampled compared to the rest of the frame (the second stratum). The relative intensity of sampling in the first versus the second stratum (r) thereby determines the gain in nominal subgroup sample size. Using proportionate sampling as the referent standard, we first compare the effect of r on the nominal and effective change in sample sizes. We then develop the optimum solution for r considering the dampening effect of variable sample weights on effective sample size (due to the varying sampling intensity between strata). Finally, as sample attrition and thus unit costs vary between strata, we also develop a solution for optimum r considering both variable weights and cost. In all findings, we take into account the impact of the correlation between the sample weights and key survey measurements and apply our results to two recent telephone surveys.

Keywords: telephone sampling; disproportionate stratified sampling; multiplicative effect of variable weights; nonresponse adjustment

1. Background

Disproportionate sampling is a commonly used tool in survey sampling when a subgroup of the population at large is of particular interest but may be inadequately represented using a proportional allocation. Carefully used, this sampling method allows for an increase in the nominal sample size of the subgroup, but it can also affect the cost of calling through the sample during data collection. Thus in many practical situations both cost and precision implications must be considered in determining optimal values of r. More specifically, the optimum value of r occurs at that value of r where the resulting precision of subgroup estimates per dollar spent (in working through the sample) is the greatest.

Figure 1: Frame/Population Size Measures

Stratum	Telephone Frame (Phone Number)		Subgroup Member? 1 (Yes)	2 (No)	Total
1 (Higher concentration)	M_1	→	N_{11}	N_{12}	$N_{1\cdot}$
2	M_2	→	N_{21}	N_{22}	$N_{2\cdot}$
Total	M		$N_{\cdot 1}$	$N_{\cdot 2}$	N

Our results here presume that the research goal is to examine phenomena in a particular population subgroup by conducting a telephone survey whose sample is chosen from two non-overlapping frames. Frame 1 is defined so that it has relatively high concentration of the subgroup, and frame 2 has a comparatively low concentration of the same subgroup (Waksberg, 1973). Furthermore, we assume that there are M units comprising a telephone sample frame that map to N unique response units in the population. Within the M phone numbers (PNs) there are M_1 and M_2 numbers on each frame, respectively, mapping to N_1 and N_2 population members (see figure 1 for details). Similarly, it can be assumed that given specific sampling rates in the high- and low-concentration strata, f_1 and f_2, respectively, these population size measures will have corresponding sample measures as seen in figure 2.

Figure 2: Sample Size Measures

Stratum	Telephone Frame (Phone Number)		Subgroup Member? 1 (Yes)	2 (No)	Total
1 (Higher sampling rate)	$m_1 = f_1 M_1$	→	n_{11}	n_{12}	$n_{1\cdot}$
2	$m_2 = f_2 M_2$	→	n_{21}	n_{22}	$n_{2\cdot}$
Total	m		$n_{\cdot 1}$	$n_{\cdot 2}$	$n \equiv n_{\cdot\cdot}$

Using these frame/population and sample size measures, we then formulate parameters which describe the relative sizes and subgroup concentrations in the two strata. Three parameters (u, v, and r) are defined to represent the relative

stratum concentrations, relative stratum sizes and relative sampling rates, respectively (see figure 3).

Figure 3: Size/Concentration Parameters

$u = \dfrac{N_{11} / N_{1\bullet}}{N_{21} / N_{2\bullet}}$ = Relative concentration of subgroup in strata

$v = N_{2\bullet} / N_{1\bullet}$ = Relative sizes of strata

$r = f_{1\bullet} / f_{2\bullet}$ = Relative sizes of sampling rates in strata (i.e., level of disproportionate sampling)

Three additional parameters reflect important practical outcomes affecting the two strata (i.e., the stratum 1 outcome divided by stratum 2 outcome): the relative stratum response rate (R_{RR}) based on response rate #4 suggested by AAPOR (2004), the relative ratio of the number of subgroup respondents to the number of assigned PNs to be called through (R_{λ}), and the relative data collection cost per assigned phone number (R_C).

Before determining an optimum value of r, it is necessary to explicitly define formulations for the variance of survey estimates and cost of calling through the sample that are relevant to a telephone survey design. In both formulations we presume that the object of survey estimates is the population subgroup. Assuming also that the sole purpose of stratification in the stratified with-replacement simple random sample (SRS) of PNs is to facilitate oversampling the subgroup, the variance of an estimated subgroup mean from a subgroup sample of size n can be expressed as

$$V = V\left(\bar{\bar{y}}\right) = Meff_{\omega^*,factors}\frac{\sigma_y^2}{n} \qquad (1)$$

where $Meff_{\omega^*,factors}$ is the multiplicative effect of variable weights linked to various study design ("factor") effects (with the superscript, *, indicating that the weights are adjusted for non-response), and σ_y^2 is the member variance of the survey outcome measure (i.e., y-variable) among subgroup members.

Two sets of factors will be considered in defining $Meff_{\omega^*,factors}$ (see figure 4). $Meff_{\omega^*,D+N}$ is the multiplicative effect of variable weights accounting for both disproportionate sampling and weight adjustment for differential non-response in the two strata, and $Meff_{\omega^*,D+N+C}$ accounts for disproportionate sampling, non-response adjustment and the correlation between the weights and the y-variable. $Meff_{\omega^*,D+N}$, is a function of the size/concentration and ratio parameters defined earlier, and the second form, $Meff_{\omega^*,D+N+C}$, can be defined as a function of the prior form with additional parameters based on the outcome measure being considered with respect to the correlation to the weights (Kalsbeek, et al, 2006). These new parameters are ρ_{y,p^*} which is the member-level correlation between the y-variable and the single-draw probability for the subgroup and α which is the intercept from a member-level regression model involving the y-variable and the single-draw selection probability for the subgroup. (Spencer 2000). It should be noted that $Meff_{\omega^*,D+N+C}$ does not equal $Meff_{\omega^*,D+N}$ under zero correlation. This inequality indicates that even if the y-variable and weights are completely uncorrelated this zero correlation still impacts the multiplicative effect of variable weights.

Figure 4: Formulations of $Meff_{\omega^*,factors}$

$$Meff_{\omega^*,D+N} = \frac{(u + rvR_{RR})(urR_{RR} + v)}{rR_{RR}(u+v)^2}$$

$$Meff_{\omega^*,D+N+C} = \theta^{(1)}Meff_{\omega^*,D+N} - \theta^{(2)},$$

where

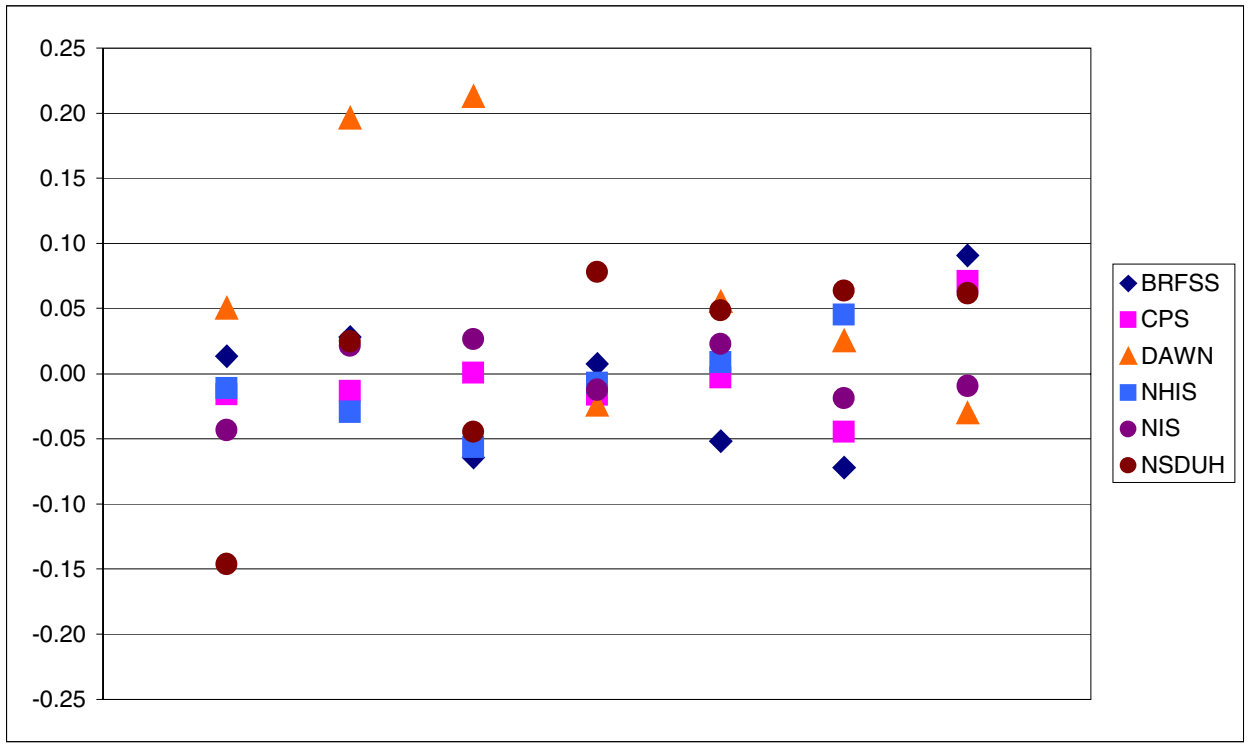
$$\theta^{(1)} = (1 - \hat{\rho}^{\,2}_{y,p^*}) + \frac{\hat{\alpha}^2}{\hat{\sigma}_y^2} \;;\quad \theta^{(2)} = \frac{\hat{\alpha}^2}{\hat{\sigma}_y^2}$$

Typically it is assumed that the correlation between the y-variables and weights is negligible and would thus have an ignorable impact on the multiplicative effect of variable weights (Kish, 1965). Given this assumption it may be sufficient to use only $Meff_{\omega^*,D+N}$ when considering the optimal sample allocation: however, this assumption may not be valid in all cases (especially considering as noted earlier $Meff_{\omega^*,D+N+C}$ does not equal $Meff_{\omega^*,D+N}$ under zero correlation).

Several different y-variables from a number of recent national surveys were used to assess the plausibility of a zero-correlation assumption and thereby the need to consider the relationship between weights and y-

variables in variance models. Each of the surveys differs somewhat in their use of disproportionate sampling methods and survey topics. As indicated in figure 5, these correlations were generally small in magnitude. Most fell within ± 0.05 of zero, which would confirm at least that a zero correlation assumption may be close to being realistic. A few correlations were found to be much larger in magnitude. It should be noted however that the weights in these studies were adjusted for non-response only and then post-stratified to further adjust for remaining sample imbalance due to nonresponse and differential frame coverage.

Figure 5: Selected values of $\rho_{y,p*}$ for an assortment of large, national surveys



BRFSS – Behavioral Risk Factor Surveillance System (2004)
CPS – Current Population Survey (March 2002)
DAWN – Drug Abuse Warning Network (1997)
NHIS – National Health Interview Survey (2003)
NIS – National Immunization Survey (2003)
NSDUH – National Survey on Drug Use and Health (2003)

Besides the variance function it is necessary to define a cost function for use in optimization. The cost of a study can be considered to consist of two parts, the first being the fixed cost of the study and the second being the variable. Fixed costs are those not affected by r, such as questionnaire development and study administration. This leaves only the variable cost to be considered in determining r. With d representing the subgroup of interest, the variable cost can be expressed as:

$$\text{Cost}_{d:Variable} = \left(\frac{rC_1 + vC_2}{r\lambda_1 + v\lambda_2} \right) n_d$$

This formulation can be simplified for computation purposes. With the C_i and λ_i parameters being those used to define two of ratio parameters defined earlier (R_C and R_λ, respectively), we can, without loss of generality, form a simpler version of the cost function which will be proportional to the original value; i.e.,

$$\text{Cost}_{d:Variable} = \left(\frac{rC_1 + vC_2}{r\lambda_1 + v\lambda_2} \right) n_d$$
$$\propto \frac{\lambda_2}{C_2} \text{Cost}_{d:Variable} \quad . \qquad (2)$$
$$\propto \frac{\left(rR_C + v\right)n}{rR_\lambda + v} = C'$$

Now that both the variance and cost functions have been defined (Equations 1 and 2), the method of optimization must be determined. Initially optimization was attempted using a method identical to that of the Neyman allocation, by which one minimizes (with respect to r) the function defined as the product of the variance, V, and cost functions, C' (e.g., see Cochran ,1977):

$$VC' = Meff_{\omega^*,factors} \frac{\sigma_y^2}{n_d} \left(\frac{rC_1 + vC_2}{r\lambda_1 + v\lambda_2} \right) n_d$$
$$= Meff_{\omega^*,factors} \sigma_y^2 \left(\frac{rC_1 + vC_2}{r\lambda_1 + v\lambda_2} \right)$$

However, it is likely that this approach would not provide an general result due to the complexity of VC', especially when V considers the correlation between the weights and y-variable. In this case a closed form solution could not be found. Given this, we decided to empirically determine optima for r using our experience with two recent telephone surveys conducted by the Survey Research Unit at the University of North Carolina.

The first survey to be used for this purpose is the Greensboro Race Reconciliation Survey (GRRS), a general population survey of adult respondents in the city of Greensboro, NC. This survey utilized a disproportionately allocated sample from a single list-assisted random digit dial (RDD) frame that was stratified by geographic area. To boost the African-American population, the subgroup of interest in the survey, geographic areas with a higher concentration of African-American households (the targeted subgroup) were sampled at a disproportionately higher rate. See table 1 for parameter values in this study.

The second survey used for illustration is the National Sun Exposure Study (NSES), which targeted the subgroup of youths aged 11-16 years old and their parents. The sample in this survey was obtained by screening households in a general population sample of all line-access PNs. Strata for the sample were formed from two different but overlapping frames of PNs, one (a demographically targeted frame of listed numbers) that was a subset of the other (a standard list-assisted RDD frame). Overlap in these previously overlapping frames had been eliminated prior to sample selection. The targeted frame was heavily oversampled in order to control survey cost. The RDD sample in this study had a much higher percentage of ineligible or inactive PNs than did the targeted frame. See table 1 for parameter values for this study.

While both of these studies utilized two-stratum samples, each was performed with different concerns. GRRS was concerned with increasing sample size for African-Americans and still being able to produce adequate estimates for the whole population, while NSES was concerned only with 11-16 year olds and their parents.

Table 1: Observed parameter values for illustrative example studies

Parameter	Study	
	Greensboro Race Reconciliation Study (GRRS)	National Sun Exposure Study (NSES)
u	2.61	4.74
v	1.39	24.84
r	1.62	21.27
R_{RR}	1.01	0.94
R_λ	1.80	16.39
R_C	0.96	2.15

2. Results

A closed form solution was found when using $Meff_{\omega^*,D+N}$ if one was to assume that R_λ and R_C were equal, thus causing the function being used for the cost reduce simply to n, the sample size. Using this new value for the cost function allowed for a closed form optimum to be determined, equal to the inverse of the ratio of the relative stratum response rates (i.e., $r_{optimal} = \frac{1}{R_{RR}}$).

This optimum is not sufficient for general purposes due to the imposed constraint. Therefore an empirical search for the optimum was undertaken. Using experience in the two previously described studies, GRRS and NSES, optimal values were determined and examined in comparison to the value of r used in each study, and in looking at the differences in the values of the optimization equation, VC'. The specific optimal values for each study were obtained

by finding the relative sampling rate (r) which minimizes the previously defined equations. This process is no longer attempting to use calculus to find an exact optimum but rather is using a more empirical, exploratory approach.

Figures 6a and 6b show graphs for both GRRS and NSES of $\frac{VC'}{\sigma_y^2}$ when only disproportionality and nonresponse adjustments are considered. Dividing the function by σ_y^2 provides two benefits; the first being to remove σ_y^2 from the function where it acts as a constant and does not have an impact on the location of the optima. The second benefit is that it allows for the optima found to be general and not specific to a single y-variable.

For GRRS (figure 6a) the actual relative sampling rate is determined when attempting to optimize considering only the effects of disproportionate sampling and ignoring the impact of both non-response and correlation between the weights and y-variable. The actual relative sampling rate (r) used for the NSES study (figure 6b) reflects an effort to minimize cost of working the survey. This rate did not include the variance in the consideration.

Notice that in Figure 6a (GRRS) the optimal and actual relative sampling rates are nearly identical. Remembering that the actual rate for GRRS was optimal considering disproportionate sampling this similarity would indicate that the additional effect of nonresponse with respect to the optimal sampling rate was minimal (the relative response rate, R_{RR}, was 1.01). Also, the difference in the functional values between the two rates is small, indicating that the impact from that difference in the optimal and actual sampling rates would have also be small.

By contrast the difference between the optimal and actual rates in Figure 6b is of a much larger magnitude, but again the difference in the functional values is small but of a greater degree than observed for GRRS. When comparing the two graphs it can be noticed that the slope of the graph for GRRS is larger than for NSES, indicating smaller deviations causing a greater impact.

Figures 7a and 7b are of a similar nature to figures 6a and 6b but also show results for optimization when correlation between the y-variable and weights is considered in addition to non-response.

Figures 7a and 7b overlay the optimization information considering disproportionate sampling, nonresponse and correlation over the results which consider only disproportionate sampling and nonresponse. Also, since correlation values are dependent on the y-variables being used in the calculations two measures were picked from each study to provide a larger picture regarding the effect caused by correlation.

Figure 6a: Optimization for the Greensboro Race Reconciliation Study considering disproportionate sampling and nonresponse.

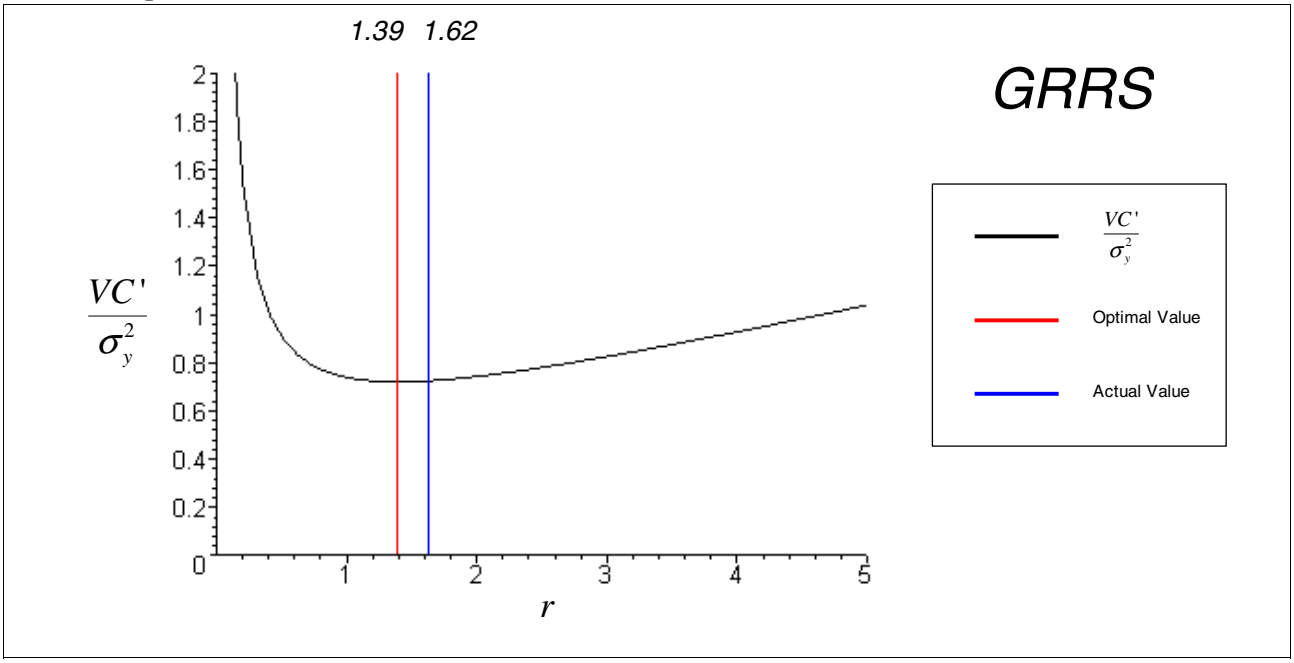
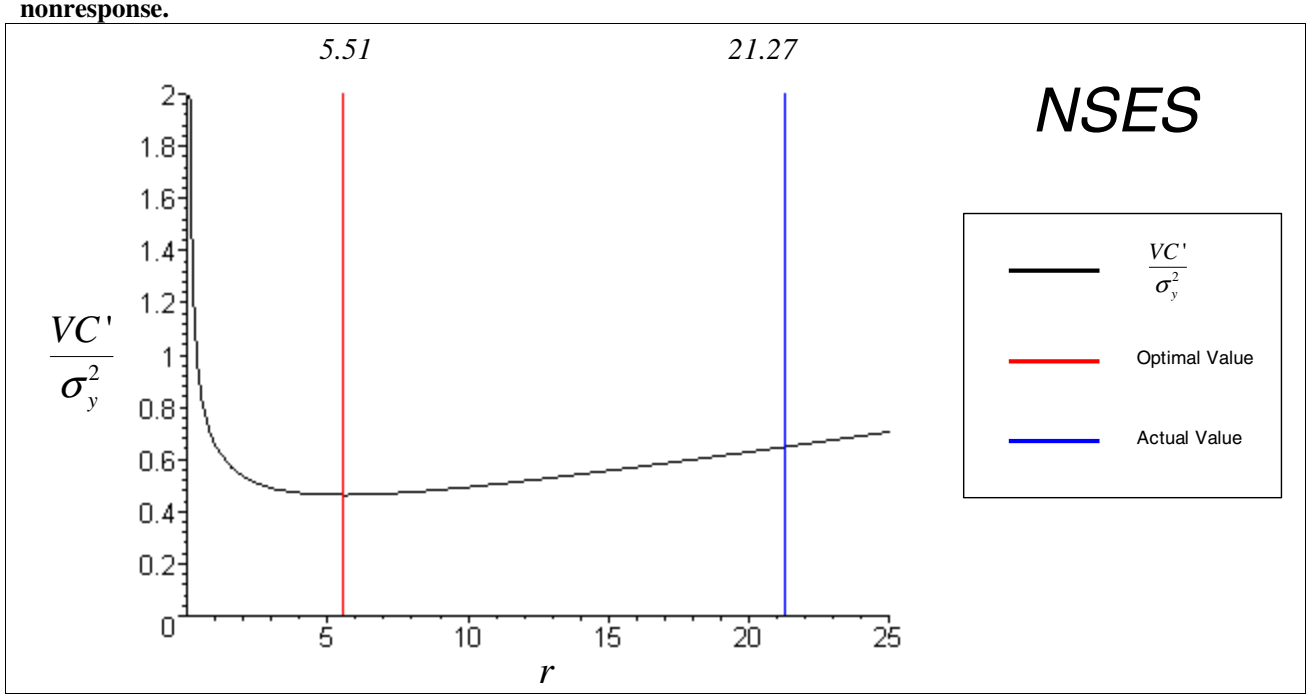


Figure 6b: Optimization for the National Sun Exposure Study considering disproportionate sampling and nonresponse.



Optimal values for the relative sampling rate for each y-variable are shown along with the actual relative sampling rate (see figures 7a and 7b). Further three functions are shown on each graph, one being $\frac{VC'_{D+N}}{\sigma_y^2}$, the function shown in graphs 6a and 6b, the other two being $\frac{VC'_{D+N+C,1}}{\sigma_y^2}$ and $\frac{VC'_{D+N+C,2}}{\sigma_y^2}$. Where the subscripts used in the above functions indicate the previously defined $Meff_{\omega^*,factors}$ functions; also the subscript 1 and 2 refer to y-variables used from each study (see Table 2 for details about the y-variables). The inclusion of the previous graph in addition to these new graphs allows for observation in how correlation affects the function being used for optimization.

For GRRS (figure 7a) the optimum for both y-variables used are virtually identical, approximately 1. Notice, the optima for y-variable 1 (r=0.99) is marginally lower than 1, possibly indicating a small amount of undersampling being optimal. This result is slightly unexpected when desiring to increase the sample size for the subgroup. Both optimization functions deviate largely from the previous function, indicating that a heavy price is paid for deviating from the optimal rate.

For NSES (figure 7b) the function for y-variable 2 deviates significantly from the previous optimization function while y-variable 1 remains similar but still indicates a price being paid for greater disproportionality. These deviations are notable since y-variable 2 which has a lowest correlation, is the NSES measure in which the greatest impact from correlation is noticed.

Table 2: Correlation related parameter values for each study and outcome measure

Greensboro Race Reconciliation Study		
Parameter	y-variable 1	y-variable 2
σ_y^2	16.431	8.498
ρ_{y,p^*}	0.006	0.031
α	18.992	28.161
National Sun Exposure Study		
Parameter	y-variable 1	y-variable 2
σ_y^2	13.658	2.548
ρ_{y,p^*}	0.088	0.001
α	2.616	3.295
GRRS: y-variable 1 – Social Cohesion Score GRRS: y-variable 2 – Racial Reconciliation Score NSES: y-variable 1 – Sunburns in a Season NSES: y-variable 2 – Sun Exposure Safety Score		

Figure 7a: Optimization for the Greensboro Race Reconciliation Study considering disproportionate sampling, non-response and correlation between the weights and outcome measure.

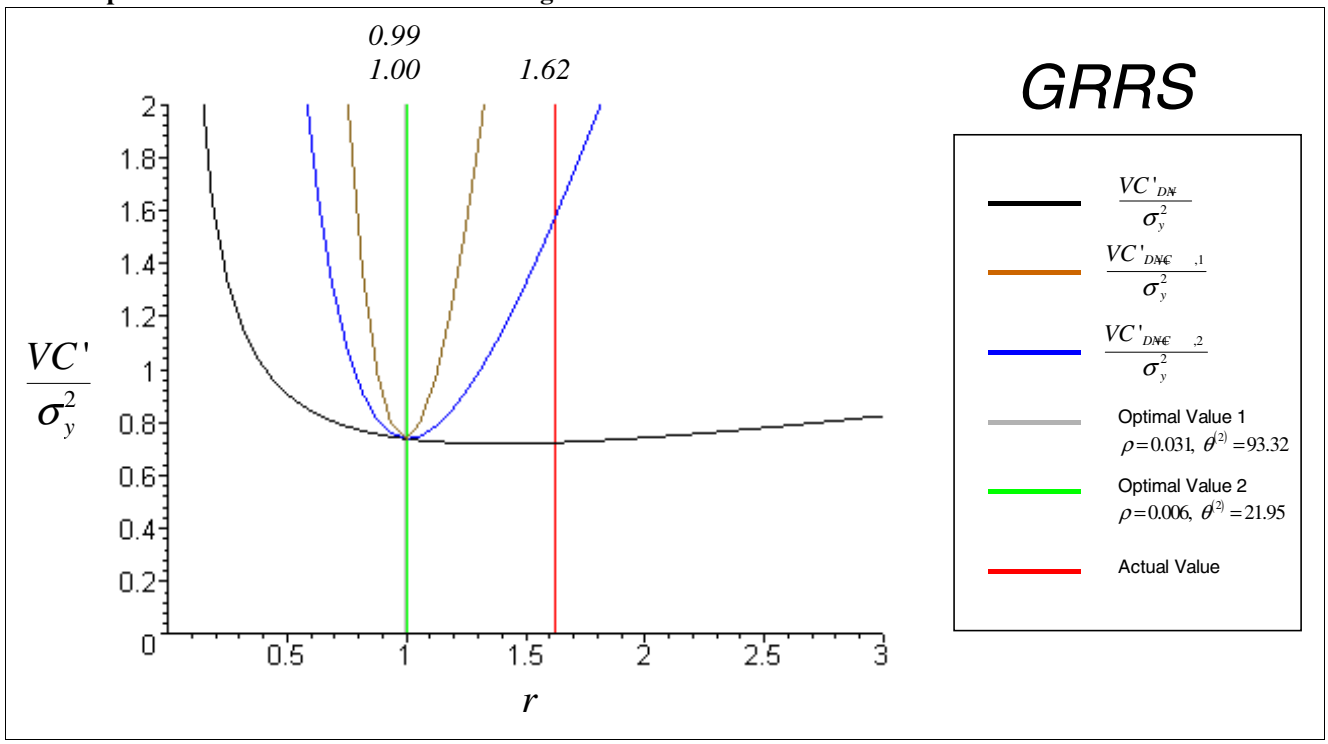
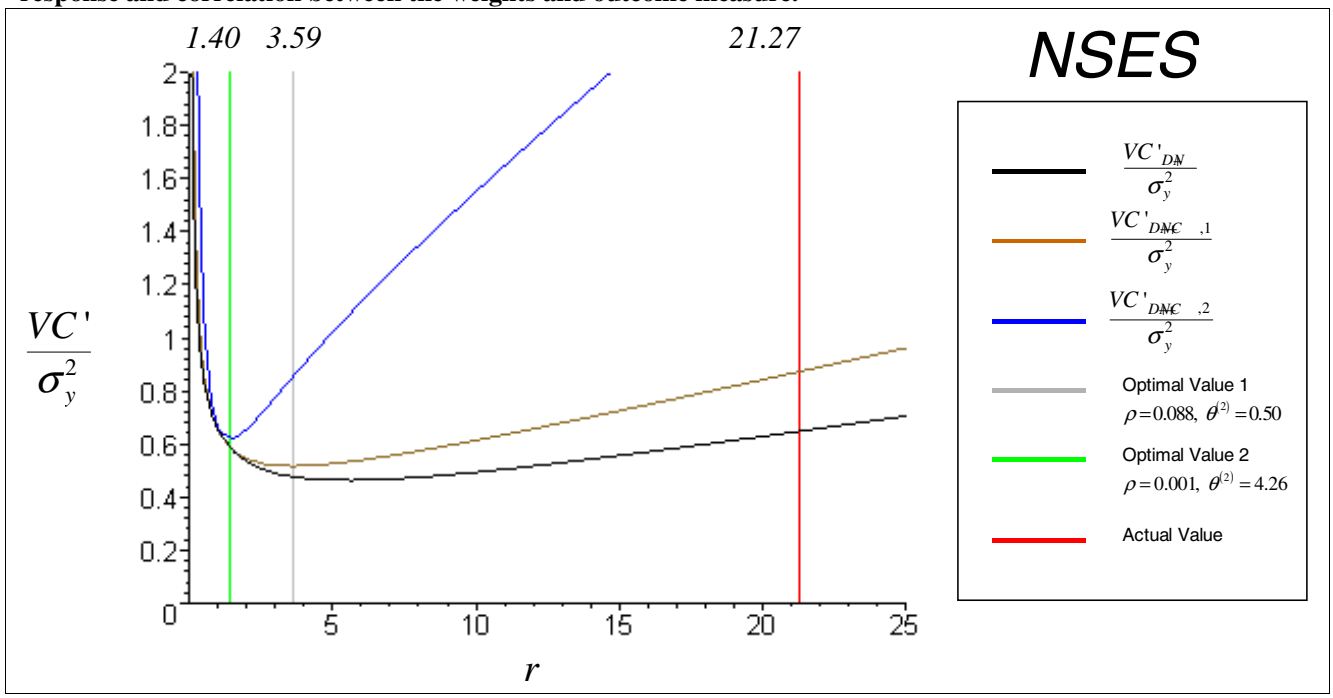


Figure 7b: Optimization for the National Sun Exposure Study considering disproportionate sampling, non-response and correlation between the weights and outcome measure.



4. Discussion

From the empirical exploration of optima based on the two examples here, one might cautiously conclude that in surveys with oversampling based on geographic stratification (such as with GRRS) the cost-efficiency of the sample will be more sensitive to deviation from the optimal values (regardless of factors being considered) than in surveys where oversampling is based on a two-frame design of the type seen in the NSES. Most curious of all is the dramatic departure in the patterns of $\frac{VC'_{D+N}}{\sigma_y^2}$ and $\frac{VC'_{D+N+C}}{\sigma_y^2}$ even when the magnitude of the correlation between the weights and the y-variable (ρ_{y,p^*}) was near zero. This suggests that the joint impact of $\frac{\hat{\alpha}^2}{\hat{\sigma}_y^2}$ and ρ_{y,p^*}, and**,** may have a greater effect on optimization than ρ_{y,p^*} alone. Thus, additional research on how the broader impact of the statistical relationship between sample weights and y-variables in surveys affects the optimum of r is needed.

References

American Association for Public Opinion Research (2004). *Standard definitions: final dispositions of case codes and outcome rates for households surveys, (3rd edition).Lenexa,Kansas.*

Spencer, B.D. (2000). "An approximate design effect for unequal weighting when measurements may correlate with selection probabilities." *Survey Methodology*26(2):137-138.

Kalsbeek, W.D., Boyle, W.R., Agans, R.P., White, J.E. (expected 2006) "Disproportionate sampling for population subgroups in telephone surveys."*Statistics inMedicine*.

Kish, L. (1965). *Survey Sampling,* New York, Wiley.

Cochran, W.G. (1977). *Sampling techniques*. (3rd Edition),NewYork,Wiley.

Waksberg, J. (1973). "The effect of stratification with differential sampling rates on attributes of subsets of the population." *Proceedings of the Survey Research Methods Section*, American Statistical Association pp. 429-434.