The Impact of Sample Design on Estimates of Demographic Behavior

Joan R. Kahn, Department of Sociology, University of Maryland, College Park
Hsiao-ye Yi, Department of Sociology, University of Maryland, College Park
Johnny Blair, Survey Research Center, University of Maryland, College Park

Key Words: Design effects; weighting; regression models; respondent selection.

The sampling specialist and the data analyst consider the sample design of a population survey from different perspectives. The sampling specialist seeks to produce the best estimates of specified population parameters given available resources, and selects from an array of available sampling strategies to do so. The sample design may include such features as unequal probabilities of selection, clustering, subsampling within households, and various post-stratification adjustments. The data analyst must decide whether to reflect these sample design features in the analysis. Two closely-related issues are whether or not to use weights for estimates, and the need to account for the sample design in computing the standard errors of those estimates. In practice, these issues are often decided based on the kind of analysis being done.

The importance of weights on estimates of descriptive population parameters such as means, totals and proportions is well known. Parameter estimators routinely incorporate weights to account for sample design features (such as unequal probabilities of selection) and for some measurement problems (such as differential nonresponse by subgroups). The sampling literature clearly demonstrates that the variances (or standard errors) of those estimators depend in part on the specific sample design, and that unequal weights often increase the variance of sample estimates (Kish 1965, p. 403).

There is far less agreement on the need or role of weights and sample design when modeling behaviors. Hoem (1989) takes the position that weighting is not necessary when one has a properly specified model since the model should control for the effects of the same factors that the weights address. Moreover, weights often bring with them unnecessary complications. In contrast, Kalton (1989), argues for the routine incorporation of weights, suggesting that "...the disagreements [over the need for weights in behavioral modeling] center on whether to rely completely on the assumptions of a carefully developed and tested model or whether to seek some protection against model misspecification." And further, "...that in most-- but not all-- circumstances it is preferable to conduct a weighted analysis and to compute standard errors appropriate for the sample design employed." Groves (1989 pp. 291-292) notes that there is no analytic resolution to this disagreement, but that it can be addressed empirically. Citing Kish and Frankel (1974), who did the first large-scale simulations of design effects on complex statistics, he notes that "clustering effects appear to influence statistics measuring both relationships and simple means." Yet, he also notes that "In every case, the design effects for means on the total sample are larger than design effects for regression coefficients for models involving those same variables."

Such empirical investigations may be useful guides for practitioners if the studies use analysis procedures and models that are discipline-specific. Most analysts of demographic behavior tend to ignore the effects of sample design. Although weights are sometimes used, other sample design features, such as clustering or within-household selection, are usually ignored. Without empirical evaluation, the impact of these decisions on a given analysis is unclear. However, it has been shown that correcting standard errors for the sample design can substantially reduce the number of significant coefficients (see Kahn, Kalsbeek and Hofferth, 1988).

In this paper, we use demographic survey data to examine the impact of sample design from the perspectives of both the data analyst and the sampling specialist. First, we consider the effects of ignoring weights and design features when estimating population parameters and regression coefficients for models of demographic behavior. We then consider the effects of alternative sample design decisions on those same estimates.

In particular, we look at the impact of conducting single or multiple interviews in the same household. While multiple interviews within households are cost efficient and allow for a self-weighting sample, they may produce response effects or less efficient estimates. For example, if early respondents discuss the survey with others in the household, they may contaminate the responses in later interviews (especially if the survey contains sensitive or personal topics). In addition, if respondents in the same household have the same or similar characteristics, this may produce high levels of intra-household correlation. Little past research has examined the impact of this type of sampling strategy on estimates of demographic behavior.

In summary, our analysis examines the impact of three design and analysis decisions: 1) whether or not to use weights, 2) whether or not to account for the sample design in computing standard errors of parameter estimates and regression coefficients, and 3) whether to conduct single or multiple interviews within households. We address these questions using actual sample survey data as well as simulations based on that data.

DATA

The analysis is based on the 1982 Puerto Rico Fertility and Family Planning Assessment Survey, an island-wide survey of reproductive-age women. The questionnaire, completed by personal interview, included a wide range of life history questions about the respondent's past educational and work experience, marriage patterns, and fertility and contraceptive behavior. The first column of Table 1 summarizes the main features of the sample design for the survey. We refer to this as the 'Original' design as distinct from the simulated sample described below.

The sample was a two-stage disproportionate stratified cluster area probability sample. The strata were SMSA and Non-SMSA based on current place of residence. Primary Sampling Units (PSUs) were clusters of, on average, 30 contiguous households. Among the two strata, a total of 150 PSUs were selected with probabilities proportionate to size (PPS) and all households within each selected PSU were sampled. Within sample households, interviews were attempted with all eligible respondents, women age 15-49. Additional weights were provided for nonresponse adjustment and for post-stratification by age. The final sample consisted of 3,175 respondents from 2,412 households.

There are several features of the survey that make the data set desirable for our purposes. First, the sample design was straightforward and intended to accommodate multi-purpose analysis (CDC, 1984); so it is reasonable to use it for both descriptive population parameters and for behavioral models. Second, the size of the clusters should produce relatively high design effects for many variables; so the decision whether to take the sample design into account should not be a trivial one. And finally, the inclusion of all eligible women in each sample household allows us to examine the effects of intra-household homogeneity. In fact, about 40% of respondents came from households with at least one other respondent. These were frequently mothers and daughters, but also sisters, roommates and other relatives.

The bottom half of Table 1 describes our analysis strategy in terms of the use of weights and within household sampling. We used the Original survey to compute both unweighted and weighted population estimates and regression coefficients. Here, we use design weights to correct for unequal probabilities of selection, nonresponse adjustment and post-stratification. We also compare standard errors computed from SAS (which ignores the sample design) with those computed using SUDAAN, which takes into account the sample design characteristics.

In order to examine the impact of conducting single or multiple interviews per household, we compared the Original data set (which interviewed all eligible women from each household) with a simulated design in which one respondent was randomly selected from each of the 2412 households. These observations were weighted by the inverse of the within-household sampling fraction as well as by the design weight. We refer to this as the '1 per household' simulation, and expect it to have smaller design effects than the Original because there is no clustering within households.

We also compare the Original sample to a second simulated design in which all respondents in a household are interviewed, but are assumed to have exactly the same characteristics. Here, we literally repeated the randomly selected observation from the first simulation until we replicated the original household size. This is clearly a hypothetical case since we would never create a sample in this way. But it provides us with an upper bound in terms of within-household homogeneity, since all observations within households are identical. Given the high degree of homogeneity, we would expect this simulated sample to have the largest design effects of all. Both simulations were replicated five times and their results averaged.

RESULTS

The analysis examines the impact on estimates of means and regression coefficients of three sample design and analysis decisions: whether to use sample weights, whether to correct for the complexity of sample design, and whether to conduct single or multiple interviews within households. We selected a small set of demographic, behavioral and attitudinal variables often used in sociological analyses and created a series of regression models loosely based on prior research. These are in no way meant to be representative of all variables or models, but rather are intended only for purposes of comparison.

Table 2a describes the variables used in the analysis and also presents the unweighted means and standard errors for the Original and Simulated samples. Recall that the main difference between the two samples is the fact that in the Original sample, all eligible respondents from each household are included, whereas in the Simulation, one respondent is randomly selected from each household. Despite this difference, we find that the mean values are very similar for the two samples on most variables. The one exception is the MARRIED variable which has a higher mean in the Simulated sample. This reflects the fact that married women are more likely than unmarried women to be the only eligible respondent from their households (since the latter tend to be younger and to live with their families). Since all households are represented in the simulated sample, married women have a higher probability of being included.

Table 2b presents weighted means, standard errors and design effects.[1] For the Original sample, two sets of standard errors are presented. The first one, calculated using SAS, ignores the features of the sample design and therefore assumes simple random sampling.[2] In contrast, the standard errors calculated using SUDAAN are corrected for the features of the sample design and, as expected, in every case are larger than the SAS standard errors. In fact, the design effects (DEFTs), which are basically the ratio of the corrected to the uncorrected standard errors, show that the majority of variables have values of at least 1.50, indicating that the corrected standard errors are at least 50% and sometimes as much as 100% larger than the uncorrected ones.

For the Simulation, we present results for the 1-per-household sample as well as the hypothetical all-per-household case in which all respondents within households are identical. Not surprisingly, given the latter sample's perfect within-household homogeneity, its DEFTs are larger than those for either of the other samples. We can think of the two simulations as providing a range for the impact of within-household clustering (i.e., since there is no clustering within households in the 1-per-household sample). When we compare the Original sample to this range, we find that in most cases, the design effects for the Original sample fall within the range for the two simulations. The Original sample DEFTs are closest to the hypothetical maximum for those characteristics which are most likely to be shared within households. These include religion (both affiliation and frequency of church attendance), education, and urban residence as a child. Variables that vary more across individuals within households (e.g., fertility, smoking, drinking), have smaller design effects, sometimes even smaller than the low end of the range for the effect of within household clustering. These results suggest that within-household clustering may have a larger effect on estimates of cultural and socioeconomic characteristics than on behaviors that reflect individual choices and preferences.

Figure 1 lists the regression models estimated. We attempted to develop models for a variety of demographic, behavioral and attitudinal variables. Logistic regression was used for the 4 dichotomous dependent variables, while ordinary least squares was used for the others. Table 3 summarizes the design effects for the models by presenting average DEFTs for all variables in each equation.

The first thing to note from Table 3 is that the design effects for the regression coefficients are much smaller than the design effects for the means. This is consistent with the findings of Groves and others. However, we still find that the design effects for the hypothetical all-per-household simulation are the largest for every model. But, rather than falling between the two Simulations, the Original sample has average design effects for every model that are smaller than even the 1-per-household simulation. This suggests that the clustering effect for this sample design may be less important in multivariate models than in simple univariate descriptive statistics.

Table 4 shows the impact of weighting and sample design correction on the number of significant regression coefficients in the models. The top half refers to the Original sample; the bottom refers to the Simulation sample where one respondent was selected from each household. The first vertical panel, which summarizes the unweighted results, breaks down the number of significant coefficients by the level of significance. The models are all relatively strong, with the majority of significant coefficients significant at the .001 level. Given the strength of these models, we are unlikely to see dramatic changes due to sampling adjustments.

As we look across the table, we see whether and how much the number of significant coefficients change after weighting (the middle panel) and after correcting the standard errors for the features of the sample design (the right panel). The symbols refer to comparisons with the unweighted results: an "=" means no change for that significance level, and a +1 or -1 means an increase or decrease by 1 significant coefficient. Despite the high levels of significance of many coefficients, we do see some changes after sampling adjustments. For the Original sample (the top panel), we find that the sample design correction using SUDAAN has a bigger effect than simply weighting using SAS. In general, there are more changes at the lower levels of significance, and in most cases, the changes lead to weakened results. For the Simulated sample, we see more changes after weighting the data than after incorporating the sample design, but this is probably due to the fact that in this sample the weights also adjust for the number of eligible respondents in the household.

SUMMARY AND CONCLUSION

In summary, we have used data from the 1982 Puerto Rico Fertility and Family Planning Assessment Survey to examine the impact of sample design on estimates of demographic behavior. Our initial objective was to look at the effect on estimates of means and regression coefficients of using sample weights, correcting standard errors for sample design, and conducting single versus multiple interviews per household.

Our results suggest that decisions about these three choices depend on the purpose of the analysis, the strength of the relationships among variables, and the type of variables under study. First, we found that design effects were considerably larger for mean estimates than for regression coefficients. This should be comforting news to data analysts who are often more interested in modeling behavior than generating point estimates. However, we also found that weaker regression results were more vulnerable than highly significant results to corrections for sample design. Hence, unless the analyst can anticipate in advance the strength of the results, it may be unwise to ignore the sample design.

Finally, we found that within-household clustering due to multiple interviews produced larger effects for some variables than others. Characteristics that are shared within households (such as religion and education) tend to be more influenced by this aspect of sample design than are behaviors that reflect individual choices and preferences (such as fertility, drinking, and smoking). From the perspective of sample design, this suggests that if the primary goal is regression modeling of variables with high expected intra-household correlations, multiple interviews within households may not be efficient. If other study objectives override this concern, and multiple interviews are conducted, it would be wise for the data analyst to consider the effects of within-household clustering on the regression coefficients.

Table 1. Description of Sample Designs and Analysis Strategy.

	ORIGINAL[a]	SIMULATION[b]
SAMPLE DESIGN		
Stratification by SMSA	Yes	Yes
# of strata	2	2
Clustering (# of PSUs)	150	150
Average Cluster Size (households)	30	30
Total # of households in Sample	2412	2412
Selection of Respondents from each household	all eligible	1 per household
Total # of Respondents	3175	2412 3175[c]

ANALYSIS STRATEGY: Alternative Uses of Weights and Within Household Sampling

ORIGINAL	SIMULATION
•No weights	•No weights
•All resp. within household	•1 resp. per household
or	or
•Design weight	•Household and design weight
•All resp. within household	•1 resp. per household
	or
	•Design weight
	•All resp. within household[c]

[a] 1982 Puerto Rico Fertility and Family Planning Assessment.

[b] Simulation created by randomly selecting one respondent from each household. This was repeated five times to create five simulated samples.

[c] Simulated sample was increased to the original sample size by replicating observations according to the original number of respondents per household. This produces maximum homogeneity within households.

Table 2a. Variable Descriptions and Unweighted Means and Standard Errors for Original and Simulated Datasets.

VARIABLE (description)	ORIGINAL Unweighted Mean	ORIGINAL Unweighted s.e.	SIMULATION Unweighted Mean	SIMULATION Unweighted s.e.
AGE (Age in Years)	30.000	0.176	30.958	0.190
CATHOLIC (Resp. is Cath.)	0.675	0.008	0.673	0.010
URBAN15 (Resp. lived in urban area at age 15) (dichotomy)	0.509	0.009	0.512	0.010
HSONLY (Resp. has HS degree but no more (dichotomy)	0.266	0.008	0.284	0.009
SOMECOLL (Resp. has completed some college but no degree (dichotomy)	0.300	0.008	0.300	0.009
GRADUATE (Resp. has graduated from college) (dichotomy)	0.104	0.005	0.118	0.007
MARRIED (Resp. is currently married or in a cohabiting union) (dichotomy)	0.559	0.009	0.643	0.010
CEB (Total number of children ever born)	1.853	0.035	1.988	0.038
CHURCH (Frequency of church attendance) (number of times per month)	2.595	0.041	2.510	0.047
ABORTION (Attitudes toward abortion (sum of positive responses to 5 yes/no questions)	1.309	0.021	1.336	0.025
SMOKE (Resp. is currently a smoker (dichotomy)	0.153	0.006	0.158	0.007
DRINK (Resp. drinks alcoholic beverages) (dichotomy)	0.229	0.007	0.235	0.009
	N=3173		N=2412	

Table 2b. Weighted Means and Standard Errors for Original and Simulated Datasets.

ORIGINAL	Weighted SAS[a] Mean	Weighted SAS[a] s.e.	SUDAAN[a] s.e.	DEFT
AGE	29.463	0.175	0.198	1.132
CATHOLIC	0.672	0.008	0.015	1.819
URBAN15	0.499	0.009	0.025	2.823
HSONLY	0.264	0.008	0.011	1.376
SOMECOLL	0.294	0.008	0.017	2.156
GRADUATE	0.102	0.005	0.010	1.830
MARRIED	0.550	0.009	0.013	1.522
CEB	1.807	0.035	0.047	1.336
CHURCH	2.557	0.041	0.070	1.704
ABORTION	1.298	0.021	0.033	1.537
SMOKE	0.153	0.006	0.009	1.368
DRINK	0.228	0.007	0.011	1.535
N=3173				

[a] SAS calculates standard errors assuming simple random sampling while SUDAAN recognizes the complexity of the sample design.

Table 2b. (continued)

SIMULATION (all using SUDAAN)

	Mean	Weighted s.e.	1 per household DEFT[c]	All per household[b] s.e.	DEFT[c]
AGE					
	29.428	0.271	1.301	0.271	1.492
CATHOLIC					
	0.671	0.016	1.669	0.016	1.914
URBAN15					
	0.494	0.025	2.537	0.025	2.911
HSONLY					
	0.263	0.013	1.380	0.013	1.583
SOMECOLL					
	0.297	0.019	2.035	0.019	2.334
GRADUATE					
	0.101	0.011	1.672	0.011	1.918
MARRIED					
	0.554	0.016	1.492	0.016	1.712
CEB					
	1.806	0.067	1.451	0.067	1.664
CHURCH					
	2.541	0.075	1.564	0.075	1.799
ABORTION					
	1.307	0.034	1.465	0.034	1.681
SMOKE					
	0.152	0.009	1.375	0.009	1.577
DRINK					
	0.237	0.013	1.560	0.013	1.790
	N=2412			N=3174	

[b]This simulated sample produces maximum homogeneity within households (see Table 1, footnote c).

[c]These DEFTs are the average across five simulated samples (see Table 1, footnote b).

Table 3 Average Design Effects (DEFTs) Across Variables within Regression Models, Estimated Using Original and Simulated Datasets.

	ORIGINAL	SIMULATION[a]	
Dependent Variable		1 per household	All per household[b]
CEB	1.135	1.195	1.366
MARRIED	1.141	1.183	1.357
HSGRAD	1.334	1.388	1.593
CHURCH	1.214	1.281	1.469
ABORTION	1.095	1.178	1.352
SMOKE	1.055	1.209	1.387
DRINK	1.059	1.158	1.328

[a]DEFTs for both simulations are averaged across five replicated samples.

[b]This simulated sample produces maximum homogeneity within households (see Table 1, footnote c).

Table 4. Changes in the Number of Significant Regression Coefficients after Weighting and Sample Design Correction.[a]

ORIGINAL

Dependent Variable	.10>p >.05	.05>p >.01	.01>p >.001	.001>p
Unweighted				
CEB	0	0	0	5
MARRIED	0	1	0	4
HSGRAD	1	0	0	3
CHURCH	0	2	0	3
ABORTION	1	1	1	5
SMOKE	1	1	0	3
DRINK	0	0	0	6
Weighted				
CEB	=	=	=	=
MARRIED	+1	-1	=	=
HSGRAD	-1	=	=	=
CHURCH	=	+1	=	=
ABORTION	=	=	=	=
SMOKE	=	=	=	=
DRINK	+1	=	=	=
Sample Design Correction (SUDAAN)				
CEB	=	=	=	=
MARRIED	=	-1	=	=
HSGRAD	-1	=	=	=
CHURCH	+2	-2	=	=
ABORTION	=	-1	+1	-1
SMOKE	=	=	=	=
DRINK	+1	=	+1	-1

SIMULATION (1 per household)

Dependent Variable	.10>p >.05	.05>p >.01	.01>p >.001	.001>p
Unweighted				
CEB	0	0	0	5
MARRIED	0	1	0	4
HSGRAD	1	0	1	2
CHURCH	0	0	2	3
ABORTION	2	1	0	5
SMOKE	1	1	1	3
DRINK	0	0	1	5
Weighted				
CEB	+1	=	=	=
MARRIED	+1	=	=	=
HSGRAD	-1	=	-1	+1
CHURCH	+2	=	=	=
ABORTION	-1	=	=	=
SMOKE	-1	+1	-1	=
DRINK	+1	=	+1	-1

Table 4. (continued)

Sample Design Correction (SUDAAN)				
CEB	=	=	=	=
MARRIED	=	=	=	=
HSGRAD	-1	+1	-1	=
CHURCH	=	+2	-2	=
ABORTION	-2	=	=	=
SMOKE	=	=	-1	=
DRINK	=	=	+2	-2

[a]Symbols reflect increases, decreases or no change (=) in the number of significant coefficients compared to unweighted results.

Figure 1. **Regression Models Estimated Using Original and Simulated Datasets.**

CEB= f (AGE, HSONLY, SOMECOLL, GRADUATE, URBAN15, CATHOLIC, MOMCEB[a])

MARRIED[b]= f (AGE, HSONLY, SOMECOLL, GRADUATE, URBAN15, CATHOLIC)

HSGRAD[b,c]= f (AGE, URBAN15, CATHOLIC, MOMCEB)

CHURCH= f (AGE, HSONLY, SOMECOLL, GRADUATE, URBAN15, CATHOLIC, MOMCEB, MARRIED, CEB)

ABORTION= f (AGE, HSONLY, SOMECOLL, GRADUATE, URBAN15, CATHOLIC, MOMCEB, MARRIED, CEB)

SMOKE[b]= f (AGE, HSONLY, SOMECOLL, GRADUATE, URBAN15, CATHOLIC, MOMCEB, MARRIED, CEB)

DRINK[b]= f (AGE, HSONLY, SOMECOLL, GRADUATE, URBAN15, CATHOLIC, MOMCEB, MARRIED, CEB)

[a]MOMCEB is the number of children ever born to the respondent's mother.
[b]Since this dependent variable is dichotomous, logistic regression is used. OLS is used for all other dependent variables.
[c]HSGRAD is a dichotomy equalling one if the respondent graduated from high school and zero
otherwise.

Acknowledgements

This research was supported in part by grants from the National Science Foundation and the William and Flora Hewlett Foundation. Computing support was provided by the University of Maryland Computer Science Center. The authors gratefully acknowledge the excellent assistance provided by Chuck McCormick, Debra Todd and Rosalind Pinkard.

References

Centers for Disease Control (1984) Puerto Rico Fertility and Family Planning Assessment (1982) Tape Contents Manual

Groves, R. M. (1989) Survey Errors and Survey Costs, Wiley

Hoem, J. (1989) "The Issue of Weights in Panel Surveys of Individual Behavior in Kasprzyk, D. et al. eds. Panel Surveys, Wiley

Kahn, J. R., W. D. Kalsbeek, and S. L. Hofferth (1988) "National Estimates of Teenage Sexual Activity: Evaluating the Comparability of Three National Surveys" Demography, Vol. 25, No. 2

Kalton, G. (1989) "Modeling Considerations: Discussion from a Survey Sampling Perspective," in Kasprzyk, D. et al. Panel Surveys, Wiley

Kish, L. (1965) Survey Sampling, Wiley

Kish, L. and M. R. Frankel (1974) "Inference From Complex Samples" Journal of the Royal Statistical Society, Series B, Vol. 36, No. 1

[1] Note that since the weight for the simulated sample adjusts for the within-household sampling fraction, there are no longer any substantial differences in weighted means between the Original and Simulation samples.

[2] This is the standard procedure used by many social scientists.