# DISCUSSION

Charles H. Alexander
U. S. Bureau of the Census
Washington, D.C. 20233*

I calculate survey weights for a living. Part of my job is answering peoples' questions about the weights. Therefore, I'm grateful to Stephanie Shipp for giving me the opportunity to review these very relevant papers about survey weights. As a test of their relevance, I've taken out my list of the most common questions I'm asked about survey weights. Let's see how well these papers answer my questions.

By far the most common question I get is:

QUESTION 1: WHEN ARE THE (Expletive Deleted) WEIGHTS GOING TO BE READY?

Stephanie asks me this at least once a month.

I'll pass over this question quickly since it is only of local interest. It does, however, point out the value of simple weighting methods. The next most common question is something like:

QUESTION 2: I'VE HEARD THAT YOU DON'T EVER NEED TO USE SURVEY WEIGHTS IF YOU (choose one):

(a) USE A SUPERPOPULATION APPROACH;
(b) ARE LOOKING AT RELATIONSHIPS OF VARIABLES;
(c) ARE USING A MODEL-BASED APPROACH.

Based on the five papers in our session, these are not sufficient criteria for ignoring weights. Three of the papers (Graubard and Korn, Bloom and Idson, and Kott) consider regression models using some version of the superpopulation approach, and each concludes that under some circumstances, survey weights should be used. Iannacchione, Milne, and Folsom assume weights will be used. Cohen and Spencer propose a procedure which always uses the survey weights, although it lets the data "decide" how much impact the weights will have.

QUESTION 3: THEN WHEN SHOULD YOU DO AN UNWEIGHTED ANALYSIS?

The three papers which consider doing an unweighted super-population regression analysis each give two answers, one theoretical and one practical.

The theoretical answers are similar:

(a) when there are "no missing regressors" (Kott)
(b) when the model is "correctly specified" (Bloom and Idson)
(c) when the weighted and unweighted estimates "are estimating the same population quantity" (Graubard and Korn).

Here "correctly specified" means not only that the variables in the model are correctly transformed and represented, but that the error term represents purely random error uncorrelated with any missing regressors.

Fortunately, the authors do not encourage analysts to simply assert their model is correctly specified. Instead, following DuMouchel and Duncan (1983) these three papers suggest testing the null hypothesis that the model is correctly specified. If the hypothesis is accepted, then proceed as though the null hypothesis is established to be true. Assuming the null hypothesis is exactly true, no weights are needed.

This is fallacious. Failing to reject the null hypothesis does not necessarily prove it is true, or even likely (or probable) to be true. Indeed for large human populations, it is hard to imagine that any model will be perfectly specified. Kott recognizes this by recommending a significance level "considerably" higher than the standard 0.1 or 0.05, because of a concern for robustness.

This really isn't a hypothesis testing problem. A more logical criterion is that weights should not be used when they cause an increase in variance which exceeds the bias reduction, i.e., when the weighted estimate has greater MSE than the unweighted. Cohen and Spencer use this criterion. Indeed, they go further and develop an estimator with lower MSE than either the regular weighted or unweighted estimators.

Practically speaking, though, the DuMouchel-Duncan test may be a reasonable procedure according to the MSE criterion. Thinking of it as a t-test, the numerator is a measure of the bias reduction from using weights. The denominator tends to be large when the weighted estimator has a high variance. Thus (roughly speaking) the larger the bias reduction compared to the variance, the larger the test statistic. The problem is where the cutoff should be. (Judging from the simple example of estimating the superpopulation mean when there are two strata with different sampling rates, Kott's advice seems good. A simple derivation shows that the weighted estimator in this simple example has the better MSE when the difference in stratum means exceeds the standard deviation of the estimated difference. This corresponds to a "t-ratio" of one, or a two-sided significance level of about 0.32).

All this assumes that there is one survey and one analysis and no need to tie the numbers in the analysis to official benchmarks or population controls. Are we really to use weighted results for some parts of a report and unweighted results for others? Do we run a stepwise regression with weights and then drop the weights once the model fits well? More important, how do we compare and consolidate unweighted estimates from different surveys whose sample designs affect the results differently? For multipurpose surveys, the simplest policy is to always use weighted estimates, as long as they have adequate precision. When weights cause unacceptably high variance, it may be better to omit them, but the best answer probably is to truncate or shrink the weights rather than drop them entirely.

Bloom and Idson try to get around these concerns by insisting on a single empirical conclusion on the importance of weights for an entire set of related analyses from different surveys. Looking at their results, it seems to me that their blanket conclusion that the weights are "unimportant" ignores some fairly important changes in magnitude for a few parameter estimates. (Several of these are not "close enough for government work." Even we professional weighters don't expect unweighted analyses to very often get the wrong sign for an important regression coefficient.) Since the weights have little effect on the authors' standard errors, their data show that they would be better off using weighted estimates throughout their

analysis to avoid these scattered instances of biased estimates.

For the papers in our session, the issue of whether or not to use weights comes back to a variance/bias tradeoff. There is, therefore, no single answer for all circumstances. A person's general position will be determined by the circumstances he/she most often encounters. My recommendation is based on dealing with multiple-purpose surveys which will often be compared with other data sources:

(a) If you can live with the variance, or if dropping the weights doesn't reduce the variance enough to help, use the weights.

(b) Otherwise shrink, truncate, or trim the weights if you can figure out a reasonable compromise solution.

(c) If not (a) and not (b), then you could use the DuMouchel-Duncan test with a significance level of 0.32.

QUESTION 4: WHAT ABOUT THE CLAIM THAT WEIGHTS SHOULD BE USED FOR "DESCRIPTION," BUT NOT FOR "ANALYSIS"?

There are more fundamentally anti-weight views than our five papers represent. DuMouchel and Duncan (1983) give a very sensible statement of one position. They note that weights may be needed when estimating the parameters of poorly specified superpopulation models. However, they present the position that an analyst should never stop with a poorly specified model, but should continue to add variables to the model until weights no longer affect the parameter estimates significantly.

Even an ardent weighter like me can't get too upset about a policy which would never deviate "significantly" from the weighted estimates, but I don't think this is a reasonable policy. For large samples, where small effects are significant, the policy would often require analysts to add many minor variables to the model. Letting weights do the "fine tuning" would permit a more parsimonious model. The focus on specifying the model to eliminate the effect of weights may distract the analyst from other kinds of misspecification. Perhaps a more likely outcome is that analysts will not bother saturating their model sufficiently to eliminate all the biases due to the sample design, but will still ignore the weights.

Pfeffermann and Smith (1985) show that for some simple weighting schemes, a careful modification of the regression model can take the sample design (plus nonresponse, frame problems, etc.) into account without explicitly using weights. This requires considerable thought on the analyst's part, and may not be a practical replacement for the many stages of weighting used on most large household surveys. However, their approach could be advantageous when the main feature of the design is heavy oversampling of a few analytically relevant strata. Little and Rubin (1987) discuss a variety of these techniques.

I have no quarrel with these modelling methods for specific analyses for which the models apply. The problem is that the existence of such methods has started the rumor that "weights aren't needed for analysis," which as it spreads becomes understood as "discard the weights and run the data through your favorite statistical package." That isn't what these papers imply.

Hoem (1989) makes arguments similar to the above, and also suggests a more extreme "no weighting" perspective. The sample may be viewed as a study population unto itself without reference to any larger population (either finite or super-). Then the relationships in the data may be analyzed (or described) based on a model which requires no sampling weights. The point is that people can do useful research without a nationally representative sample. The problem is that the alternative to a nationally representative sample should be a well-defined study population. A household survey sample, analyzed without its weights, is a heterogenous and ill-defined study group. This makes scientific inferences difficult to generalize. Especially it would be hard to compare results from diferent surveys which have different but also ill-defined, study groups.

These more fundamentally anti-weight arguments are asserted to apply to "analysis;" weights are to be used when the purpose is "description."

I think it is an unfortunate dichotomy between "analysis" which need not be based on accurate description, and "description" which has no need for analytic value. Also, the terms are not used consistently. Pfeffermann and Smith (1985) use "description" only for finite population estimation, while Cohen and Spencer seem to include superpopulation estimates without a correctly specified model.

QUESTION 5: WHERE DID YOU GET YOUR RULES FOR "COLLAPSING" OR TRUNCATING WEIGHTS?

Household survey weights are usually truncated (or weighting classes are "collapsed") when they get too large, in hopes of approximating optimal MSE for the most important survey characteristics. This partially addresses the objections to using weights. Typical truncation rules have little systematic theory behind them. Although in its early stages, the work of Cohen and Spencer is a step in this direction. They cite an impressive advantage of their method over those of Stokes (1990) and others, but that debate will continue.

QUESTION 6: HOW PROMISING IS THE METHOD OF WEIGHTING FOR NONRESPONSE BY USING LOGISTIC REGRESSION TO ESTIMATE RESPONSE PROPENSITY?

The household survey weighting community has been buzzing about the idea of deriving noninterview weights from a logistic regression model for the probability of nonresponse, at least since its mention in Little (1986). Of all the ideas in this session, it may be the most likely to have a fundamental near-term effect on how survey organizations calculate their weights.

Iannacchione, Milne, and Folsom use this method to adjust for unit nonresponse in a follow-up survey of soldiers' spouses. Besides applying the method, the authors give a clear demonstration that it preserves certain weighted marginal totals from the full sample. In this respect it is similar to an alternative "raking" method. My suspicion is that the two methods give similar results. However, the logistic regression method does a better job of incorporating continuous explanatory variables, and seems to be easier to use.

The paper also introduces an informative "Receiver Operating Characteristic" curve method, borrowed from another field, for testing the significance of a two-stage nonresponse adjustment when the stages are not independent. The test is overkill for their survey, but the authors make

good use of the ROC curve to describe differences between military paygrades.

The authors' application of the logistic regression method is sound. The weights depend on the choice of variables for the model, but so do weights in the traditional weighting class method. In the hands of skilled and careful model builders, reducing the problem to a familiar regression model selection task may reduce the arbitrariness of this model dependency. I can see only two concerns:

(a) The paper does not mention any check to be sure that the model does not lead to very large weights for a few cases, which could increase the variance excessively.

(b) There is some danger in having "completeness of the last section of the Soldier Questionnaire" be the primary predictor of the probability that a soldier will provide his/her spouse's address. This may have the effect of giving higher weight to respondents who give erroneous data. This may still be a good idea, but the data need to be well edited.

I wish the paper had said more about the general merits of the method, compared to traditional methods. There are several theoretical grounds for using the method, and I'd like to know which they would choose.

The Spouse Survey is an atypical situation because so many explanatory variables are available for a spouse whose soldier has responded. The method is particularly attractive in this situation. Logistic regression lets the authors use as many variables as they want; the regular weighting class method would have constrained the number of variables they could use.

The regression method is less essential for surveys which have few explanatory variables for nonrespondents, or which have higher response rates. Indeed, the authors do not use the method to adjust for initial nonresponse by the soldiers. Even in this more typical case, I see potential advantages of the logistic regression method:

(a) In many applications, the weights would vary less than weights from the weighting class method with a large number of cells.

(b) Once the model has been fit, the weights are easier to calculate than the traditional method which requires collapsing of weighting classes. However, for ongoing surveys, there is an issue of how often the model needs to be re-estimated.

The main practical drawback of the regression method is the careful data analysis needed to derive the model. This may rival the effort expended to analyze the weighted data. Given this level of effort, it may become attractive to analyze the response mechanism jointly with the main survey variables.

QUESTION 7: IF YOU USE WEIGHTED DATA, DON'T YOU NEED SPECIAL FORMULAS TO ESTIMATE THE VARIANCE?

A common misperception is that if you ignored the weights, you could analyze survey data as a simple random sample. Unfortunately, the effect of the sample design has to be taken into account with or without the weights. This is recognized in all these papers.

The Graubard and Korn paper which was presented here is part of a larger and very substantial paper, dealing with methods for incorporating a complex sample design into a superpopulation regression analysis. The paper includes an illuminating discussion of alternative methods for testing hypotheses about the regression coefficients. The authors investigate asymptotic and finite-sample significance levels and power with a well chosen set of simulated normally distributed data. The discussion gives intuitive reasons for many of the simulation results.

Graubard and Korn give some useful tentative recommendations on when to use which testing procedure. The best procedure depends on the relationship of the number of parameters in the regression model and the number of strata used in the variance calculation. The Wald procedure, which requires estimation of a covariance matrix, is preferable when the number of strata (and hence the number of degrees of freedom) is large compared to the number of parameters. When the number of strata is small, so the whole covariance matrix can't be estimated well, use either a version of the Rao-Scott procedure (based on the estimated eigenvalues of a matrix related to the covariance matrix) or a version of the jackknife procedure suggested by Fay. The numerical results indicate the Fay procedure for means and the Rao-Scott procedure for regression coefficients. The paper did not give an intuitive explanation of the latter result; the proofs in the appendix concerning the jackknife hint that perhaps its convergence to the asymptotic results is slower when dealing with nonlinear estimates like regression coefficients.

I think these five papers did a pretty good job answering my questions. They have a lot to add to the ongoing discussion about when and how to use survey weights. Although I think some of the authors were too eager to ignore weights, all the papers are useful because the authors clearly state their assumptions, procedures, and results.

* The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau. The author thanks Sandy Davis for her assistance in preparing this manuscript.

References

DuMouchel, W. H. and Duncan, G. J. (1983). "Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples," Journal of the American Statistical Association, 78, 535-543.

Pfeffermann, D. and Smith, T. M. F. (1985). "Regression Models for Grouped Populations in Cross-Section Surveys," International Statistical Review, 53, 37-59.

Hoem, J. (1989). "The Issue of Weights in Panel Surveys of Individual Behavior," in Panel Surveys, ed. D. K. Kasprzyk, G. Duncan, G. Kalton, M. P. Singh: John Wiley & Sons, New York, 539-565.

Little, R. J. A. (1986). "Survey Nonresponse Adjustment," International Statistical Review, 54, 139-157.

Little, R. J. A., and Rubin, D. B. (1987). Statistical Analysis with Missing Data: John Wiley & Sons, New York.

Stokes, L. (1990). "A Comparison of Truncation and Shrinking of Sample Weights," Proceedings of the Bureau of the Census 1990 Annual Research Conference: Bureau of the Census, Washington, D.C., 463-471.