

# **To Weight or Not to Weight - That is the Question**

**William G. Cumberland  
Department of Biostatistics  
UCLA**

# Introduction

“Survey weighting is a mess. It is not always clear how to use weights in estimating anything more complicated than a simple mean and standard errors are tricky even with simple weighted means.” Gelman (2007)

# History of Survey Weights

Horvitz and Thompson published their seminal paper in 1952, showing how weighting inversely to selection probability led to unbiased population estimates:

$$E_{\mathcal{P}}\left(\sum_{i \in S} \frac{y_i}{\pi_i}\right) = T$$

Here, expectation  $E_{\mathcal{P}}$  is taken with respect to the Randomization Distribution - which is created by the selection process, and  $\pi_i$  is the selection probability.

For population characteristics, weighting by inverse probability of selection gives intuitively appealing estimates.

# Problems with the Theory

Horvitz-Thompson also provided an unbiased (with respect to  $\mathcal{P}$ ) variance estimate, which frequently gave negative results. Yates and Grundy were quick to point this out.

Basu questioned the whole process of inverse-probability weighting in a cleverly constructed example of the “Circus Statistician.”

So being right on average does not necessarily mean one is making good inferences.

# Basu's Example

A circus owner wants to ship 50 elephants, but needs to estimate the total weight. He plans to choose the average elephant, Sam, weigh it and multiply by 50. The circus statistician is horrified and insists on a random sampling plan, but in a compromise with the owner uses pps sampling:  $\pi_{sam} = 99/100$  and  $\pi_j = 1/4900$  for  $j \neq sam$ .

Of course Sam is selected, but now using H-T estimator, the estimate for the weight of the herd is  $y_{sam}/.99$ . The owner, now incredulous asks what the estimate would be if the plan had selected Jumbo the largest elephant. The unhappy statistician answers:  $y_{jumbo} \times 4900$ .

This is how the Circus Statistician lost his job and became a professor of statistics.

# Basu Explained

Of course this is an example of inappropriate use of statistics. H-T estimates are optimal when  $y_j/\pi_j$  are nearly constant. Here the sampling plan is arbitrary, and the  $\pi$ 's bear no relation to the  $y$ 's. An appropriate model here would have gotten around this problem.

Blind application of weighted estimates leads in this case to an estimator which is right on average (unbiased), but never right in any particular implementation.

# Model-based Inference

In the late 1960's, cracks began to appear in the reliance on the Randomization Distribution for finite population inference.

Godambe noted that the likelihood function with respect to  $\mathcal{P}$  was flat - sanctioning any inference for the population that was consistent with the data in the sample.

Royall questioned the validity of inference based solely on  $\mathcal{P}$ , with his “ass, axe, box of old horseshoes” example, arguing that if the measurements in our population cannot be sensibly modeled, we cannot make sensible inferences about them.

# The Modelers

Inference was reformulated as a prediction problem, using a model to link the measurements. Different prediction techniques - BLU's, likelihood, Bayesian - were introduced.

In these, the sampling plan, and consequently the sampling weights played no direct role, unless introduced into the model as covariates.

The fundamental issue was robustness of the inferences to model misspecification.

# Descriptive versus Analytic Surveys

Survey sampling has always made a distinction between estimates of population characteristics such as totals and means (Descriptive Surveys), and estimates of relationships between measurements on subjects, such as regression coefficients (Analytic Surveys).

On top of this, there were the issues of design-based inference versus model-based inference.

For estimates of means, proportions, and totals, the use of weights generally leads to sensible common-sense estimates. Model-based approaches often lead to basically the same estimates.

# Regression Estimates

For estimating a total, weighting by probability of selection has considerable appeal and face-validity. If someone is selected to be in the sample with probability 1/1000, then intuitively that person represents 1000 people. Indeed,  $\sum_{i \in S} w_i \sim N$ , and equals the population size in expectation.

But for regression, the picture is far less clear about the use of weights in estimating slopes and relationships. Should an odds ratio for disease risk as a function of race be weighted because African Americans in the sample were oversampled?

# Model Misspecification

Using an incorrect model is a problem. The estimates of regression coefficients are sensitive to which units are sampled, and weighted estimates can be different from unweighted ones.

All models are wrong, some are just more wrong than others (my apologies to Orwell).

So we need to always be concerned about using a model which reasonably describes our data distribution.

# Design Properties

Survey weights are not in fact just based on probability of selection. They start out this way, but are adjusted for non-response, and are further adjusted to match population characteristics in post-stratified ratio adjustments and/or raking.

Modelers promote the ideal that the mechanism by which the data were collected is irrelevant to the inferential process. This is true for both proponents of a Bayesian approach and those who advocate likelihood prediction.

# Ignorable Designs

Rubin and others have pointed out that such analyses implicitly assume that the sampling design is ignorable. If the design is to be ignorable, then the analysis must include anything which affects the probability a person is included in the sample. This means all measurements that could influence sampling or nonresponse should be included in any regressions if they could have anything to do with the outcome. And this means they should be included not only as main effects but also in interaction terms.

This is a tall order, and a number of simulations showed that many modeling attempts fell short of this ideal.

# Asymptotic Design Properties

Pfeffermann, among others, considered the role that weighted regression estimators might have. He argued that including the sampling weights directly in the regression estimation (which most statistical packages now easily accommodate) has several advantages:

1. Such estimators have some protection against biases that result from model misspecification
2. P-weighted estimators are under “mild” conditions, asymptotically design unbiased and consistent for their targets, even when the sampling design is not ignorable.

# Problem Solved?

So if P-weighted estimates have these desirable properties, why not just always use them? The reasons are both philosophical and practical.

First, the use of P-weighted regression can entail very large losses in efficiency when the weights vary a lot. This may not be of much concern for major outcomes in large surveys, but for secondary analyses, it can lead to loss of significance for important relations.

# Loss of Efficiency due to Weighting

The (over)simplified formula for loss of efficiency is

$$\frac{(\sum w_i)^2}{n \sum w_i^2}$$
 which is always less than 1, unless all weights

are equal. When the weights vary a lot, this loss can be dramatic.

In the CHIS, weights can vary tremendously for some subgroups which are sampled from two sampling frames. The loss can be so extreme that a weighted estimate using additional observations from the second frame barely changes the standard error.

# Principles of Inference

Inferences should be made using models - this is one of the basic principles of statistics. Models should be flexible enough to incorporate features of the design, and if these are essential to inference, the model will help explain what their effect is.

Relying on asymptotic properties of the sampling design, hoping it will make up for inadequate modeling of the data is a poor approach.

# Other Problems with Weighted Inference

Asymptotic normality is not easy to show with weighted estimates, making small sample distributions of P-weighted estimates difficult to determine.

Hence many standard procedures of inference are not applicable with P-weighted estimates. These include likelihood ratio tests, and residuals analysis.

# What to do?

Using both approaches can help choose the appropriate model for inference from a sample. If the model being used and P-weighted estimates differ substantially, this suggests that the design is informative and more care should be used in formulating the model. These models may need to contain interaction terms between design characteristics and other predictors in the model. Once an appropriate model is chosen, the estimates should be made without resorting to weights. This approach has the added advantage of easily incorporating random effects into multi-stage designs that accommodate dependence of observations within clusters.

## References:

Basu D. (1971), Foundations of Statistical Inference

Gelman A. (2007), Stat. Sci. 22:153-164

Godambe V.P. (1966), JRSS-B 28:310-319

Horvitz D. and Thompson D. (1952), JASA 47:663-685

Korn E. and Graubard B. (1995) JRSS-A 158:263-295

Pfeffermann D. (1996), Stat. Methods Med. Res 5:239-261

Royall R.(1968), JASA 63: 1269-1279

Rubin D. (1976), Biometrika 63:581-592