

Survey Design and Weighting

(material drawn from Korn & Graubard 1999)

- A. Some basic design considerations and their implications
 - 1. Many considerations when conducting a survey
 - a. want the final results to be representative of an analysis population
 - b. may want the survey to include specific subpopulations
 - i. racial/ethnic diversity
 - ii. economic diversity
 - iii. age diversity
 - iv. programmatic diversity
 - c. want to keep costs down by surveying many people in a limited number of areas
 - d. may have other design issues, e.g., AddHealth samples within schools and also samples peer networks
 - e. possible differential unit-non-response
 - 2. These design issues affect subsequent statistical analyses
 - a. affect the representativeness of the observations
 - b. also affect the independence of the observations
 - 3. Consider some general issues
 - a. clustering, selecting subjects from a few areas, may lead to spatial correlation in the data
 - i. although this might not lead to biased estimates of population parameters, it could affect calculations of standard errors
 - ii. essentially, there is less variation than a sample taken completely at random

- b. oversampling of particular populations, which is done to ensure adequate representation of those individual populations, leads to a sample that doesn't represent the general population
 - c. different rates of unit non-response can also lead to a loss in representativeness
4. In analyses, we address these issues by
- a. including survey weights
 - b. including design variables

B. Sampling plans

1. Simple random sampling
 - a. consider a given population of size N
 - b. choose a subset of n individuals where each possible subset is equally likely to be sampled
 - c. individuals are chosen without replacement (we won't refer to this distinction subsequently)
 - d. the ratio n/N is called the sampling rate or inclusion probability
 - e. most sample estimators, such as the usual mean and variance estimators, assume this type of sampling
2. Stratified simple random sampling
 - a. population is first divided into mutually exclusive and exhaustive strata
 - b. simple random sampling is then carried out *within each strata*
 - c. sampling rates may (will likely) vary across strata
 - i. for example, if the populations of the strata differ but the sample sizes don't, the sampling rates will

- vary
 - ii. sampling rates may vary for other reasons
 - d. subsequent population estimators would weight each observation by the inverse of its sampling rate
 - e. if the observations are more homogeneous within strata than across them, stratified random sampling can reduce the variance of the population estimates
3. multi-stage sampling
- a. population is first divided into cells, or *primary sampling units* (PSUs), usually on the basis of geography
 - b. a sample of those units is taken
 - c. sampling of individuals then takes place within those selected PSUs
 - d. advantages
 - i. reduces the costs of conducting a survey by restricting it to a smaller number of areas (very important if in-person interviews are used)
 - ii. also may be the only feasible way to construct a sampling frame
 - e. disadvantage is that observations within clusters may be correlated
 - f. weighting is again used to adjust estimates; weights are the product of
 - i. inverse of the PSU inclusion probability and
 - ii. sampling rate within the PSU
 - g. here we've described a two-stage process, but the process can be carried out iteratively
4. Unit non-response
- a. this is a distinct issue from the sampling plan;

- however, the correction would be similar
- b. within PSUs and clusters, we may (will) have unit non-response
 - c. estimated response probabilities can be formed, possibly conditioning on additional information
 - d. the inverse of these probabilities can then be multiplied times the other components of the weight calculation to form a new weight
5. In all of the cases, except simple random sampling, we have unequal sampling rates for individuals within the overall population but can use weights to obtain unbiased estimates of the population parameters

C. Poststratification

1. Another strategy to improve the accuracy and representativeness of a sample is to *poststratify* the sample or the sampling weights
2. In poststratification, weights are developed so that the totals or proportions of different types of respondents match “known” population figures
3. The advantage of this technique is that it brings in additional information about the population
 - a. it can be effective in dealing with differential non-response and under-sampling
 - b. it also helps to make surveys more comparable by reweighting them to a common analysis population

D. Use of weights in statistical analyses

1. If observations are sampled unequally, respond

unequally, or can be adjusted to reflect the larger population, it makes sense to use weights to reduce biases

2. Weights are usually the product of
 - a. inverse sampling probabilities, sometimes referred to as the base weight (note: these sampling probabilities can themselves be products of probabilities if multi-stage sampling is used),
 - b. inverse response probabilities, sometimes referred to as non-response adjustments, and
 - c. poststratification adjustments
3. For most cross-sectional statistics, calculations for weighted estimates are straightforward
 - a. let X_i be a variable for subject i and let W_i be the associated weight
 - b. the weighted mean is

$$\bar{X}_w = \frac{1}{\sum W_i} \sum W_i X_i$$

- c. let x_i and y_i , be deviations of variables X_i and Y_i from their weighted means; the slope coefficient from a weighted regression of Y_i on X_i is

$$\beta_w = \frac{\sum W_i x_i y_i}{\sum W_i x_i^2}$$

4. In SAS, most statistical procedures include a WEIGHT statement
 - a. syntax

WEIGHT <weight_variable>;

- b. the *weight_variable* would be the SAS variable containing the weights
5. If the sampling design indicates that weights should be included, you should include them, right?

E. An alternative to weighting

1. An alternative to weighting is to model the survey design in your statistical procedure
2. In a multivariate model, this is accomplished by including measures of the characteristics that enter the weighting procedure as additional explanatory variables in an unweighted model
 - a. coefficients on these variables will confound genuine effects and survey design effects
 - b. however, coefficients on the other variables should be purged of the influence of design effects
 - c. this assumes that you have modeled the design correctly
 - d. it also assumes that the design variables don't introduce other problems
3. Weights are often based on characteristics that we would include in models anyway, such as age, race/ethnicity, age, and socioeconomic status
4. Suggests that weights might not be especially useful in multivariate analyses
5. Other considerations might lead us to drop weights
 - a. dropping incomplete cases (cases with item non-response) from a weighted sample changes the

response pattern in that sample

- i. formally, the sample should be reweighted to reflect the new sample
 - ii. however, this is rarely done
 - iii. result is incorrect weights that might not reduce bias
- b. in some cases, weights can increase, rather than reduce, the variance of estimators; we might consider bias vs. efficiency (MSE) trade-offs

F. Adjusting for clustering

1. If a clustered survey design is used, observations may not be independent within clusters
 - a. this can lead to incorrect (usually downward biased) standard errors
 - b. it also means that the estimation procedure is inefficient
 - c. in the simplest cases, the problem is similar to that from a random effects specification
 - d. issues become more complicated if weights and other design issues enter
2. Two types of corrections are possible
 - a. just fix the standard errors—this leads to correct standard errors but does not address the efficiency concerns
 - b. estimate a FGLS specification
 - i. addresses standard errors and efficiency
 - ii. however, it requires you to take a stand regarding the precise source of spatial correlation
 - c. most researchers simply choose to address the first

problem, using a robust method for calculating standard errors

3. A practical difficulty in adjusting for clustering is that not all public-use surveys include the necessary identifiers

G. SAS procedures

1. As mentioned, most SAS statistical procedures allow you to incorporate WEIGHTs
2. SAS has a number of procedures that are designed to accommodate survey data with complex designs
3. PROC SURVEYMEANS¹
 - a. syntax

```
PROC SURVEYMEANS <options>;  
  VAR <list_of_analysis_variables>;  
  STRATA <list_of_stratifying_variables>;  
  CLUSTER <list_of_clustering_variables>;  
  WEIGHT <weight_variable>;
```

- b. the sample design information is specified in the STRATA, CLUSTER, and WEIGHT statements
- c. if multiple STRATA or CLUSTER variables are specified, SURVEYMEANS examines the available combinations of these variables
- d. if STRATA and CLUSTER variables are both specified, SURVEYMEANS adjusts for clustering within STRATA
- e. for multi-stage designs the first (highest) STRATA

¹ See http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/surveymeans_toc.htm.

- and CLUSTERS should be used
- f. the procedure allows for other options such as CLASS analyses and BY processing

4. PROC SURVEYREG²

- a. syntax

```
PROC SURVEYREG <options>;  
  STRATA <list_of_stratifying_variables>;  
  CLUSTER <list_of_clustering_variables>;  
  WEIGHT <weight_variable>;  
  MODEL <dependent_variable> =  
        <list_of_independent_variables>  
        </ options>;
```

- b. the sample design information is specified in the same way as the SURVEYMEANS procedure
 - c. unlike the standard REG procedure, only one MODEL statement can be specified
- #### 5. “SURVEY” procedures are also available for
- a. frequency distributions—SURVEYFREQ
 - b. logistic regression—SURVEYLOGISTIC
 - c. selecting samples with given design features—SURVEYSELECT

² See http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/surveyreg_toc.htm.

References:

Carrington, William J., John L. Eltinge and Kristin McCue. "An Economist's Primer on Survey Samples." Working Paper no. 00-15. Suitland, MD: Center for Economic Studies, U.S. Bureau of the Census, October 2000.

Korn, Edward L., and Barry I. Graubard. *Analysis of Health Surveys*. New York: Wiley, 1999.