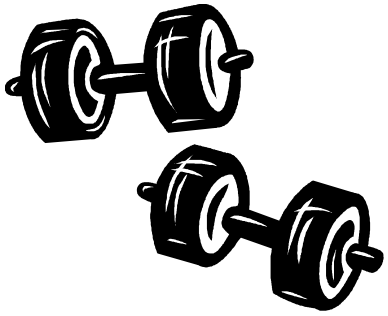

Using Weights in the Analysis of Primary and Secondary Data

Anastasia R. Snyder (snyder.893@osu.edu)

Department of Human Sciences

College of Education and Human Ecology

The Ohio State University

Goals for Today's Talk

- Discuss the pros and cons of using large secondary data sets with complex sampling design.
- Understand how complex sampling affects data.
- Provide an overview of weighting to correct for design effects related to complex sampling.
- Demonstrate how to ***svyset*** your data using Stata
- Cover the complications associated with weighting
- Provide an example of using weights in a primary data collection project.

What is a Survey Weight?

- A value assigned to each case in the data file.
- Normally used to make statistics computed from the data more representative of the population.
- The weight value indicates how much each case will count in a statistical procedure.
- Examples:
  - A weight of 2 means that the case counts in the dataset as two identical cases.
  - A weight of 1 means that the case only counts as one case in the dataset (unweighted).
  - Weights can (and often are) fractions, but are always positive and non-zero.
- [in Stata, these are the *pweights*]

What is a Survey Weight?

- Two most common types:
  - **Design Weights**—To accommodate design effects created by complex sampling
    - stratified sampling, multistage cluster sampling
  - **Post-Stratification** or **Non-response weights**—compensate for non-response among individuals with certain traits.
  - **Can and often do use both**

Types of Survey Weights

- **Design Weight**: (over or under sampling and clustering effects)
  - Normally used to compensate for over- or under-sampling of specific cases or for disproportionate stratification.
  - ***Example*:** It is a common practice to over-sample minority group members. If we doubled the size of our minority sample, then each minority case would get a design weight of ½ or .5
  - The design weight when we want the statistics to be representative of the population.

Post-Stratification Weights

- **Post-Stratification** or **Non-response** Weight.
  - This type is used to compensate for the fact that persons with certain characteristics are not as likely to respond to the survey.
  - *Example*. Most general population surveys have substantially more female than male respondents (often 60/40). Because the survey over-represents females and under-represents males in the population a weight is used to compensate for this bias.
  - There are many respondent characteristics that are likely to be related to the propensity to respond.
    - Age, Education, Race/ethnicity, Gender, Place of residence

How Do We Calculate Weights?

- For analysis, only one weight per case can be used. If we weight for different factors, these weights must be combined together into one weight.

- Lets say we have a design weight (Dwate) and a post-stratification (PSwate) weight for each case.

- To calculate a total weight these are multiplied together:
- Total Weight = Dwate * Pswate

- *Note*: never give a weight the value of 0 unless you want the case excluded from the analysis. It should default to 1.

Calculating Design Weights

- If we know the sampling fraction for each case, the weight is the inverse of the sampling fraction.

- Design Weight = 1/(sampling fraction)

- The sampling fraction could also be the over-sampling amount for a given group or area.

- *Example:* If we oversampled African Americans at a rate 4 times greater than Whites, than the design weight for an African American would be ¼ (.25) and for Whites would be 1.

Calculating Post-Stratification Weights or Non-response Weights

- This is typically more difficult then design weights.

- It requires the use of auxiliary information about the population and may take a number of different variables into account.

- Information usually needed:
  - Population estimates of the distribution of a set of demographic characteristics that have also been measured in the sample
  - For example, information found in the Census such as:
    - Gender, Age, Educational attainment, Residence (e.g., rural, urban, metropolitan), Region

Sources for Auxiliary Statistics for calculating Post-Stratification weights

- Population data for community-based samples:
  - U.S. Census tabulations
  - The Current Population Survey (CPS)
  - Department of Education

- For other types of surveys source can be:
  - Reports or enrollment data from a school or university.
  - Organizational statistics data are from an organization.

- Finding good estimates for the population characteristics is sometimes a challenge.

Calculating Post-Stratification Weights

Gender	Population Proportion	Sample Proportion	Population/Sample	Weight
Female	.5	.6	.5 /.6	.8333
Male	.5	.4	.5 /.4	1.25
Total	1	1		

Census report is used to find the gender distribution in the population (50% female). This is compared to the gender distribution in the sample of completed interviews (60% female.

Problem: *What if you have more than one characteristic to balance with the population?*

Adjusting for Multiple Population Characteristics

- **Options for combining characteristics:**
  - You can combine characteristics in a single table to do the calculation:
    - Males 18-25
    - Males 26-45
    - Females 18-25
    - Females 26-45
- **However**:
  - You need to have these crosstab tables available for the population source
  - The number of cases in each cell in the sample cannot be too small.

- **Therefore:** It may be better to use several separate frequency tables rather than one big N-way crosstab to compute the weights, especially when several characteristics are being balanced.

Calculating Post-Stratification Weights when you use separate frequency tables

- **Example:** You have separate tables for the age, gender, education, race/ethnicity, metropolitan status for the population. [these are not crosstabed with each other]

- Single variable frequency tables are more likely to be available for the population.

- Use of frequency tables may reduce unstable weights due to small Ns in the sample that may occur if comparing N-way crosstabs.

- The challenge is how do you combine the weights for each characteristic?

Calculating Post-Stratification Weights

- Different options for combining the weights.
  - 1. Compute a weight for each characteristic independently and then multiply all these weights together.

    NOT RECOMMENDED. Will usually not yield good weights.
  - 2. Compute weights separately but sequentially.
    - Calculate a gender weight comparing the population and sample gender distributions.
    - Weight the sample data by the gender weight.
    - Generate the frequency distribution for education after the data are weighted by gender.
    - Calculate the education weight.
    - Weight the data by gender and Education (multiplying the weights) and generate the weighted Age (in categories) frequency distribution.
    - Calculate the age weight.
    - Etc.

Problems with these approaches

- This second approach is better, but the characteristics early in the sequence are not likely to match the population when the later characteristics are adjusted.
    - The gender percentages may not be the same in the sample and population after the education and age weights are included in the total weight.
    - This can occur when the characteristics may be correlated (e.g. Age and education)
- Several possible solutions to this problem.

Three Possible Solutions

- 1. Use a single big age x gender X education table for the calculation of the weights (BEST SOLUTION).
  - However, crosstabs may not be available for the population
  - and, small cell sizes in the sample table

- 2. Iterative Solutions:
  - Manual version (stepwise programming in statistical software)
  - Automatic version (i.e. Raking software)

- 3. Logistic regression based solutions if case level population data is available.

Manual Iterative Solution

- Example with three characteristics A, S, E
  - 1. Compute A weight (wA) and weight data by this weight
    - Generate the weighted frequency table for S
  - 2. Compute S weight (wS) and weight by wA*wS
    - Generate the weighted frequency table for E
  - 3. Compute E weight (wE) and weight by wA*wS*wE
    - Generate the weighted frequency for A
  - 4. Compute a second A weight(wA2) and weight by wA*wS*wE*wA'
    - Generate the weighted frequency for S
  - 5. Compute a second S weight (wS2) and weight by wA*wS*wE*wA2*wS2
    - Generate the weighted frequency for E
  - 6. Compute a second E weight (wE2) and weight by wA*wS*wE*wA2*wS2*wE2
  - Continue process until the weighted frequencies and the population frequencies don't change. Usually converge after two or three iterations (or less)

Automatic Iterative Solutions

- A procedure, called **Raking**, has been programmed by several folks. Is relatively widely used.
- There is a Raking ado for Stata.
- In the SAS macro you can set several options, such as how accurate you want to weight, and also can impose some limits on the size of weights (min and max).
- The SAS Raking macro is pretty clunky and hard to use.
- The Stata ado has fewer options.

Logistic Regression Approach to Weighting

- This approach requires that you have a dataset that you are using for the population figures (e.g. the PUMS data, CPS, or ACS datasets)
- **Example:** CPS Public Use data set for 2014 includes age, education, race (in categories), gender, and metropolitan status variables.
  - Assume you have the same variables measured in the same way in the data set you want to weight to increase representativeness.
  - Create a subset of the CPS with just these variables and add an indicator called "Sample" set equal to 0. Also create of subset from your survey with the same variables formatted the same as the CPS data, but set the Sample" equal to 1.
  - Combine the cases from the two data sets together. (Stack data versus merge)
  - Use "sample" as a dependent variable in a logistic regression with each of the other characteristics as independent variables. Set the regression program to save the predicted probability (pprob) from the regression for each case and include it in the dataset.
  - The weight would be the inverse of this predicted probability. (Weight = 1/pprob)
  - Yields weights that are highly correlated with those obtained in raking.

  - EXAMPLE OF HOW I'VE CALCULATED WEIGHTS FOR RYE STUDY

Problems with Weights

- Weights primarily adjust means and proportions. OK for descriptive data but may adversely affect inferential data and standard errors.
- Weights almost always increase the standard errors of your estimates. Introduce instability into your data.
- Very large weights (or very small ones) can also introduce instabilities.

Problems with Weights

- Some researchers like to “trim” the weights. To not allow extremely high weights that can increase instability of estimates.

- Trimming the weights can often result in reducing the representativeness of the weighted data.

- Trade off between less instability or more accurate representativeness.

- Several techniques have been developed to try to reduce extremes in the size of the weights and still yield representative results.
  - Collapsing categories
  - Putting constrains in the iterative process on the relative size of weights (e.g., found in the SAS Raking macro).
  - Various Bayesian and MCMC methods have been developed to yield more stable weights, so far have not been used much. This is beyond my knowledge.

Data Analysis Methods with Weighted Data

- Should use a statistical procedure that adjusts for the impact of the weights on the standard errors. Standard errors based on the actual N and not the weighted N.
  - Not available in SPSS. SPSS treats weights incorrectly in inferential statistics
  - SVY procedures in Stata.
    - Also use of pweight. Weights in SAS normally treated correctly.

- Normalization of weights.
  - Setting the weights so the N in the weighted data equals the N in the unweighted data.
  - To calculate, multiply the weight by (Unweighted N)/ (Weighted N)
  - If the statistical procedure does not use weights correctly for the standard errors, normalization is a less biased choice.

- Another choice is to not use weights at all for regression models. Instead include all the variables used to create the weights as independent (control) variables. Results in unbiased estimates and standard errors.

Household vs. Individual Level Weights

- Many datasets have both a household and an individual level weight (Census, CPS).

- Use of household vs. individual weights.
  - Interview surveys are often sampled and conducted at the household level.
  - One respondent, usually at random, is selected to be interviewed.
  - The weight needs to take into consideration the differential selection of individuals in households
    - For household with only one adult the sampling fraction is 1/1
    - For household with 3 adults the fraction is 1/3
    - Unless weighted (as inverse of the sampling fraction) a bias towards single adult household results.

- Use household weight when you want to generalize to characteristics of households (like poverty rate)

- Use individual (person) weight when generalizing to a population of individuals

What Weights to use in Analysis of Longitudinal (Panel) Data?

- Many panel data sets have several weights to choose among.
  - Cross-sectional weights (first wave weight)
  - Weights for each panel if multiple panels

- Weights to use will primarily depend on the data analysis methods used.

- Longitudinal Panel weights are usually computed from two components
  - 1. The **cross-sectional** weight from the previous panel or the first panel
  - 2. A weight calculated to adjust for **attrition** between the waves.

- Calculating the non-response (attrition) weight component:
  - Usually use logistic regression with response to the wave as outcome variable (attrition 0= no; 1=yes).
  - Predict probability of responding (use logistic regression to predict attrition)
  - Inverse of this probability is the attrition weight (if prob=.883, attrition wt=1/.883).

What Weights to use in Analysis of Longitudinal (Panel) Data?

- **Example 1**: Four-wave panel
  - Waves in 1997, 2000, 2003, 2006.
  - Plan to analyze the respondents to the 2003 wave, but use data from 2000 and 1997 as well. Maybe with a growth curve model.
  - Should use the panel weight for 2003.

- **Example 2**: Same data as above
  - Plan to analyze all four-waves using a random or fixed effects model.
  - All respondents in each wave are retained in the analysis.
  - Should use the 1997 cross-sectional weights.

- **Principle**:
  - If respondents in the analysis are those from a specific panel, then use the weights for that panel.
  - If you want to follow respondents from a specific wave forward, then you should use the weights for that specific wave.

When to use Unweighted Data

- If the sample is not self-weighted then it is a good idea to use weights as often as possible.

- Some methods don't allow weights.

- Steps to follow to avoid bias in unweighted analyses:
  - Include as independent variables in the models (control for) all the variables that might account for the disproportionate sample design or non-response.
  - If a weight is available, the weight itself could also be included as an independent variable.
  - If the weight has a significant effect on the outcome in a model including the design variables, then it suggests the weight is likely to have been constructed in a way related to the dependent variable. A bias is possible.
  - Compare weighted and unweighted results from methods that allow weights. If no substantive differences, then weights yield a bias.
  - Weighting has a larger effect on descriptive statistics then on regression coefficients.

Summary

- Most statistical software programs allows for weights and most treats them properly.
  - Stata is considered the best program for large survey data with complex sampling designs.

- If you have specific questions about using weights, or calculating weights for your primary data collection project, please feel free to contact me and I will try to answer them if I can.