Sampling Weights and Variance Estimation

Jerry West
Senior Fellow
Overview

- Introduce you to
  - Sampling weights
  - Methods for calculating variances and standard errors for complex sample designs
- General introduction to these topics
- Weights are unique to research studies and data sets
- Options for calculating variances and standard errors will vary by study
Webinar Goals

- You will have a basic understanding of sampling weights and variance calculations for complex sample designs that are used in many of the national early care and education studies.
- You will use weights and use procedures that are appropriate for calculating variances and standard errors for these designs in your research.
- You will increase your chances of having your work published by doing these things.
Sampling Weights
Sampling Weights: What’s All the Fuss About?

- Generate more questions from data users than any other topic
- Use and misuse in professional literature

Today’s topics
- What are sampling weights and why are they important?
- Why use weights in your analyses?
- Does using weights make a difference?
- Why so many different weights?
- Can I use weights with my software? How?
- Frequently asked questions on weights, when and how to use them?
What is a Sampling Weight?

- A sampling weight is used to indicate the relative strength of an observation.

- In the simplest case, each observation is counted equally.

- For example, if we wish to calculate the mean age for webinar participants, we just sum everyone’s age and divide by the number of participants.
What is a Sampling Weight (continued)?

- When we use unweighted data, each observation is counted equally.

- Unweighted data represents only the observations of those in the sample that provide data.

- When we weight the data, the observations are counted relative to sample members’ representation in the population.

- Weights allow analyses that represent the target population.
Why are Weights Important?

You can get different answers:

<table>
<thead>
<tr>
<th>Value</th>
<th>4</th>
<th>2</th>
<th>1</th>
<th>5</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

- **Simple mean**: \((4+2+1+5+2)/5 = 2.8\)
- **Weighted mean**:
  
  \[
  \frac{(4\times1) + (2\times2) + (1\times4) + (5\times1) + (2\times2)}{10} = 2.1
  \]
  
  or
  
  \[
  \frac{(4+2+2+1+1+1+1+5+2+2)}{10} = 2.1
  \]
Sample Designs for Early Care and Education Studies

- Many are designed to produce national estimates (FACES, ECLS-K, ECLS-B, HSIS, NHES)
- Rely on sample data, i.e. data on the entire population was not collected
- Programs, schools, teachers, classrooms, and children did not have an equal probability of selection.
- Not all selected programs, schools, teachers, classrooms, parents, children participated.
Why Use Weights?

- Weights compensate for not collecting data from the entire population and for using a complex sample design

- Weights
  - Adjust for differential selection probabilities
  - Adjust for differential nonresponse and attrition
  - Reduce potential bias associated with nonresponse
  - Are used when estimating characteristics of the population
### Does Using Weights Make a Difference? (ECLS-K)

<table>
<thead>
<tr>
<th>Base Year Characteristic (C1CWO)</th>
<th>Unweighted</th>
<th>Weighted (Base Weight)</th>
<th>Weighted (Final Weight, C1CWO)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>57</td>
<td>56</td>
<td>58</td>
</tr>
<tr>
<td>Black</td>
<td>15</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Hispanic</td>
<td>18</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Asian</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>School Type</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td>78</td>
<td>87</td>
<td>85</td>
</tr>
<tr>
<td>Private</td>
<td>22</td>
<td>13</td>
<td>15</td>
</tr>
</tbody>
</table>

Source: ECLS-K
## Does Using Weights Make a Difference? (HSIS)

<table>
<thead>
<tr>
<th>Child Outcome</th>
<th>Unweighted</th>
<th></th>
<th>Weighted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beta T vs. C</td>
<td>SE Beta</td>
<td>Beta T vs. C</td>
<td>SE Beta</td>
</tr>
<tr>
<td>PPVT</td>
<td>3.93*</td>
<td>1.47</td>
<td>4.01*</td>
<td>1.98</td>
</tr>
<tr>
<td>WJ Applied Problems</td>
<td>3.19*</td>
<td>1.21</td>
<td>2.83</td>
<td>2.16</td>
</tr>
<tr>
<td>WJ Word</td>
<td>5.94*</td>
<td>1.3</td>
<td>5.76*</td>
<td>2.87</td>
</tr>
</tbody>
</table>

* p <= .05
Why So Many Different Weights?

- Most data sets for the large-scale national studies have many different weights.

- Weights vary according to:
  - Level of analysis: child, teacher/classroom or program/school
  - Round(s) of data: cross-sectional or longitudinal
  - Source(s) of data: child assessment, parent interview, and/or teacher questionnaires

- Choice of weight to use will be driven by your research question.
There are 43 weights on the FACES 2006 data files, how do I know which one to use?

- Weight should match your research question
- Consider level of analysis, sources of data, and whether your question requires one or more than one round of data
- There is no perfect weight!
### SUMMARY OF DATA COLLECTION COMPONENTS, BY WAVE

<table>
<thead>
<tr>
<th></th>
<th>3-Year-Olds</th>
<th>4-Year-Olds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fall 2006</td>
<td>Spring 2007</td>
</tr>
<tr>
<td><strong>3-Year-Olds</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Parent Interview</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Teacher Child Rating</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Teacher Interview</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Classroom Observation</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Center Director Interview</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Education Coordinator Interview</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>4-Year-Olds</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Parent Interview</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Teacher Child Rating</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Teacher Interview</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Classroom Observation</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Center Director Interview</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Education Coordinator Interview</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
When children enter Head Start, do their language, literacy, and early math skills differ by gender and race/ethnicity?

- How many rounds of data are needed?
- What is/are the source(s) of these data?
- What is the level of analysis?
Example 2 – Choosing the Best Weight

- Do the gains children make in their language, literacy, and math skills across their first year in Head Start differ by their skill level at program entry?

  - How many rounds of data are needed?
  - What is/are the source(s) of these data?
  - What is the level of analysis?
Example 3 – Choosing the Best Weight

- Are the gains children make in their language, literacy, and math skills during their time in Head Start influenced by the quality of their Head Start classrooms and teachers?
  - How many rounds of data are needed?
  - What is/are the source(s) of these data?
  - What is the level of analysis?
How to Use Weights with Commonly Used Software Packages

- **Use**
  - “WEIGHT” statement in SAS, or
  - “WEIGHT BY xxxxx” in SPSS
  - [weight=xxxxx] in Stata

- **Weights often sum up to population totals.**

- **Normalizing weights**
  - Normalize weights for certain software packages (SPSS)
  - Impact greatest on estimates of standard errors and significance tests
How do I Normalize the Weights?

- Calculate a new weight that sums to the sample size
- The new weight is equal to the weight found on the data file multiplied by the ratio of the sample size (number of cases with a positive value for the given weight) to the sum of the weights.
ECLS-K Example - Normalizing Weights

- Weight to be normalized: C2PW0
- Sum of weights: 3,865,946
- Total number of cases with a positive weight: 18,950
- Normalized weight = C2PW0 * (18,950 / 3,865,946)
- Sum of normalized weight = 18,950
Other Frequently Asked Questions

- When selecting a weight, do I have to subset my dataset?

- What happens to cases where there is no positive weight?

- What weights do I use if analyzing a subsample of cases?

- What if I’m running a regression - what weights do I use?

- What weight do I use if I’m using a multi-level model?
Summary

- Weights should be used when analyzing data from national studies of early care and education such as FACES, ECLS-K, ECLS-B, HSIS, and NHES.

- The appropriate weight should be selected based on:
  - Level of analysis
  - Round(s) of data
  - Source(s) of data

- Weights adjust for unit, but not for item nonresponse

- There may not be a “perfect” weight for some analyses. The best weight can be determined with some descriptive analyses.
Participant Questions
VARIANCE
Overview: Calculating Standard Errors

- Why are variances and standard errors important?

- Why not use standard errors that assume a simple random sample (SRS)?

- How to use “design-based” methods for estimating standard errors.

- How to use approximation methods for estimating standard errors.
Why are variances and standard errors important?

- Data from national early childhood and education studies are used to estimate population characteristics and to test hypotheses
  - differences between two or more groups of children, classrooms, schools/programs
  - relationships between characteristics of children and/or their environments

- Standard errors are a measure of the variability of those estimates

- Standard errors are used when testing whether the differences or the relationships are statistically significant
Most standard procedures in commonly-used statistical software packages assume data are from a simple random sample.

The procedures are not appropriate for complex multi-stage, clustered sample designs with unequal probabilities of selection.

Software packages designed for simple random samples tend to underestimate the standard errors for complex sample designs.
<table>
<thead>
<tr>
<th>Children’s Reading Score</th>
<th>Mean T-Score (Fall K)</th>
<th>Standard Error</th>
<th>Standard Method (SRS)</th>
<th>Design-Based (JK2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All children</td>
<td>50.1</td>
<td>.080</td>
<td>.215</td>
<td></td>
</tr>
<tr>
<td>Below poverty</td>
<td>44.5</td>
<td>.163</td>
<td>.316</td>
<td></td>
</tr>
<tr>
<td>At or above poverty</td>
<td>51.4</td>
<td>.087</td>
<td>.189</td>
<td></td>
</tr>
</tbody>
</table>
Getting the Right Variance

- Design-based variance estimation methods
  - Replication methods
  - Taylor Series methods
- Approximation method
- Design-based methods are preferred
Design-Based Methods

- Replication techniques
- Replicate weights
  - ECLS-K, ECLS-B, and HSIS replication weights use jackknife methods.
  - Software to use
    - WESVAR replication series (JK2, JKn)
    - AM (JK2)
    - SUDAAN (JK2)
Design-Based Methods

- Taylor Series
  - Use PSU and strata IDs from data file
  - Software
    - SUDAAN
    - STATA
    - SAS (using PROC SURVEY)
    - AM
Approximation Method

Two steps:

1) Normalize weights
   - May need to adjust weights so standard error is based on actual sample size rather than population size

2) Use design effect (DEFF)
   - To account for complex sampling design
Step 1 - Normalizing Weights

- Weights on FACES, ECLS-K, ECLS-B, and HSIS sum to the population totals.
- Calculate a new weight that sums to the sample size
- The new (normalized) weight is equal to:
  
  \[ \text{sampling weight} \times \frac{n(\text{weight})}{N(\text{weight})} \]
ECLS-K Example - Normalizing Weights

- Weight to be normalized: C2PW0
- Sum of weights: 3,865,946
- Total number of cases with a positive weight: 18,950

Normalized weight = 
\[ C2PW0 \times \left( \frac{18,950}{3,865,946} \right) \]

Sum of normalized weight = 18,950
Step 2 - Adjusting for Complex Design

- Use design effect (DEFF) to adjust SRS standard errors

- What is a design effect?
  
  - Ratio of the variance of dependent variable produced by specialized software that accounts for the complex design to the variance of the same dependent variable produced using procedures that assume SRS.
Using design effects

\[
\text{DEFF} = \frac{\text{Design variance}}{\text{SRS Variance}}
\]

\[
\text{DEFT} = \sqrt{\text{DEFF}} = \frac{\text{Design standard error}}{\text{SRS standard error}}
\]

Example for fall ECLS-K kindergarten reading scores

SE (SRS) = .063
SE (Design) = .156

\[
\text{DEFF} = \frac{.156^2}{.063^2} = \frac{6.15}{6.15} = 6.15
\]

\[
\text{DEFT} = \frac{.156}{.063} = \sqrt{6.15} = 2.48
\]
3 Ways of Using the DEFF

- Multiply the SRS standard error produced by the square root of the DEFF (i.e., DEFT).
- Adjust the t-statistic by dividing it by the square root of the design effect (DEFT) or adjust the F-statistic by dividing it by the DEFF.
- Adjust the sampling weight such that an adjusted standard error is produced.
Using a DEFF-Adjusted Weight

- First step, create a weight that sums to the sample size (normalized weight).
- Second step, divide this normalized weight by the DEFF.
- Use this weight for analyses and the standards errors produced will approximate the standard errors obtained using “exact” methods.
For SAS Users

- SAS base procedures such as PROC REG, PROC FREQ, PROC MEANS do account for the actual sample size but not for complex sampling.

- SAS procedures such as SURVEYMEAN and SURVEYREG (procedures that begin with "SURVEY"), use Taylor series method to provide exact estimates by incorporate complex sampling procedures into the standard error estimation.
Regression Analysis with Complex Sample Designs

- Use appropriate software such as AM, WESVAR, SUDAAN, or SAS (SURVEYREG procedure).

- For SAS (PROC REG procedure) use DEFF-adjusted weights.

- For SPSS, use normalized, DEFF-adjusted weights.
Summary

- National early childhood and education studies use complex sample designs.
- The appropriate weight should be used depending on the type of analysis.
- The standard error calculation should be based on sample total rather than population total.
- The standard errors should be adjusted for the complex design either using software or the design effects.
Participant Questions