GSA Webinar Series

HRS Sample Design, Weighting, and Complex Variance Estimation

DEVELOPED AND PRESENTED BY THE UNIVERSITY OF MICHIGAN WITH FUNDS FROM THE NATIONAL INSTITUTE ON AGING

Webinar Panel

Amanda Sonnega, PhD
Associate Research Scientist
Survey Research Center
Institute for Social Research
University of Michigan

Sunghee Lee, PhD
Associate Research Scientist
Survey Methodology Program
University of Michigan

Ryan McCammon, MS
Research Area Specialist Lead
University of Michigan
Quick Audience Poll

- Please tell us who you are:
  - Graduate Student/Post-Doctoral Fellow
  - PhD Faculty
  - Dual Degree Faculty
  - Undergraduate Student
  - Other
OVERVIEW

- Provide basic understanding of key issues in the HRS sample design
  - Sample weights
  - Complex variance estimation
- Provide tools needed to address these issues in analysis of HRS data
  - Where to find the relevant variables
  - Applying the weights and design variables in analysis
  - Some advice on using weights
COMPLEX SAMPLE DESIGN

- Objective of a sample is to represent a population of interest
- Simple random samples are straightforward
  - Each sample element has equal probability of selection
  - Sample from a complete list of all elements
  - Sampling error associated with sample estimates
  - E.g., Jill Politician has 49% of votes with +/-3% margin of error
- Not always optimal or even possible
- Complex sample design is an alternative
  - Defining feature is that not all sample elements have equal probability of selection

COMPLEX SAMPLE DESIGN

- Two main techniques: stratification and clustering
  - Stratification:
    - Classify the target population into subgroups (strata) whose members are assumed to be more like each other than members of other strata
    - Sampling then done within each stratum
    - E.g., study of English dialects in the United States. You know this varies substantially by region. To ensure you capture dialects in each region, first create regional strata and sample within them
COMPLEX SAMPLE DESIGN

- Clustering:
  - The target population is divided into groups (clusters), a sample of clusters is randomly selected, then each member of the cluster is interviewed
  - E.g., study grade schoolers. There’s no list of names and addresses of all grade school students. But do have addresses of grade schools, and schools have lists of students. But it’s not realistic to obtain all of those lists. Rather, randomly select a set of schools (clusters) and interview all students in the school.
- Other design features can be added such as oversampling of certain population groups of interest

SRC NATIONAL SAMPLE FRAME

- When the entire country is the population of interest, not reasonable to do simple random sampling
  - No list of all US residents exists
  - Cost would be prohibitive to locate and interview such a dispersed sample
- Survey Research Center (SRC) at the University of Michigan’s Institute for Social Research maintains a national sample frame of dwellings for all area probability surveys
- Using it, surveys implement complex sample designs
Entire population divided and subdivided in several stages

- Form primary sampling units (PSUs), which are usually counties or county groups
- Within the primary units, blocks, or clusters of addresses in cities and towns, and chunks of rural areas are the second stage units
- In a third stage of sampling, we select small segments or clusters of dwelling units where interviews are taken for a study
- In a fourth stage, individuals are selected from the dwelling units
HRS SAMPLE DESIGN

- HRS uses a stratified multistage sampling design
- In 2016, undertook modifications for greater efficiency and changing demographics
- Cohorts through Mid Baby Boomers included oversampling of Hispanic and African American households

Stage 1
- Select PSUs with probability proportional to population size (more populous, greater chance of selection)

Stage 2a
- Select secondary sample units, SSUs, with probability proportional to size
- SSU = blocks or “rural chunks”

Stage 2b: Use US Postal Service delivery sequence list to assemble a list of all the housing units in SSU

Stage 3
- Select individual housing units with probability INVERSELY proportional to size
- Evens out the selection probabilities and improves efficiency
MULTISTAGE AREA SAMPLING

- **Stage 4a**
  - List household members (birth year and relationship) and select cohort-eligible units

- **Stage 4b**
  - If more than 1 eligible unit in the household, randomly select 1

- **Note**: institutions (prisons, hotels, nursing homes, etc.) not sampled

IMPORTANT IMPLICATIONS

- The major implication of the complex design is that sample units (both individuals and households) have differential probabilities of being selected into the sample

- Inferences to the US population will be biased unless sample weights are used in calculating descriptive statistics such as means and totals
  - Sample weights account for differential selection

- The sample is geographically stratified and clustered so variance estimates using standard formulae are not appropriate
  - Need to be corrected in analyses using weights
HRS provides household and respondent-level sample weights

- Constructed to make HRS weighted sample representative of all US households containing at least one person in the age-eligible range (in the case of household weights) or of all non-institutionalized individuals in the US population age 51 years and older (in the case of respondent weights)
WEIGHT COMPONENTS

- Baseline sample weights include:
  - Base weight (based on the sample design)
  - Inverse of the probability of selection, fixed, positive
- Non-response adjustment at each wave. Higher for:
  - Men than women
  - Single than married
  - Young than old (ignoring mortality)
  - African American and Hispanics than non-Hispanic white
- Post-stratification
  - Project the sample to target population
  - Uses an outside data source (with lower nonresponse) as a reference
  - We used the Current Population Survey (CPS)
  - Changes to HRS weights pursuant to corrections to CPS data (Ofstedal, Weir, Chen, & Wagner, 2011)
  - Now using the American Community Survey

RESPONDENT AND HOUSEHOLD WEIGHTS

- Household weights
  - For analyses at the household level
  - E.g., wealth, housing
  - Positive as long as one of the respondents is age eligible, an original respondent, living and not institutionalized
- Respondent weights
  - For analyses at the respondent level
  - E.g., health, work, cognition Positive only if the respondent is age eligible, connected to an original respondent, living and not institutionalized
COMPLEX VARIANCE ESTIMATION

WHY BE CONCERNED?

- Default on most statistical packages assumes simple random sampling (SRS)
- Multistage samples are not SRS
  - Stratification (reduces sampling variance)
  - Clustering (increases sampling variance)
  - Not self-weighting (increases sampling variance)
- We need to correct the standard errors in analysis using two design variables, STRATUM and SECU (more in a moment)
DESIGN EFFECTS

- Design effect = ratio of sampling variance under complex design to sampling variance under a simple random sampling design
- Represents the effect that the complex sample design has on the variance
- Measure of efficiency of the complex design

IMPORTANCE OF DESIGN EFFECTS

- Design effects are in relation to specific variables (no overall design effect for a survey)
- Design effects will be highest for variables that are:
  - Homogeneous within clusters
  - Strongly related to weights
TAKEN TOGETHER...

- Sample weights affect point estimates and may influence variances and standard errors
- Adjustment for complex sample design affects only variances and standard errors

WHERE TO FIND THE RELEVANT VARIABLES
WEIGHT VARIABLES IN HRS

- HRS provides wave-specific weights
  - Both household and respondent level
- Available in Tracker file (in background docs)
  - XWGTHH (household)
  - XWGTR (respondent)

- $X$ ranges from $A=1992$ wave to $P=2016$ wave

- 2016 weights will be available soon

DESIGN VARIABLES IN HRS

- STRATUM ranges from 1 to 56
  - Defines the sampling error computation strata of the HRS data
  - 2 strata for New York, Los Angeles, and Chicago
  - Note that the values of STRATUM will be different starting with the 2016 wave because of changes to the sampling strategy
- SECU (code 1 and 2)
  - The stratum half-sample code for analysis of sampling error using the balanced repeated replication method or approximate “two PSUs-per-stratum” Taylor Series method (Kish & Hess, 1959)
  - Always need 2 PSUs per stratum for the analysis to work
APPLYING WEIGHTS AND DESIGN VARIABLES IN ANALYSIS

SVY COMMAND IN STATA

```
.svyset strata stratum
.svysset psu secu
.svysset pweight FMGTR
.svydes

pweight: FMGTR
Strata: stratum
PSU: secu
```

<table>
<thead>
<tr>
<th>Strata</th>
<th>#PSUs</th>
<th>#Obs</th>
<th>min</th>
<th>mean</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>347</td>
<td>145</td>
<td>173.5</td>
<td>202</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>347</td>
<td>158</td>
<td>173.5</td>
<td>189</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>212</td>
<td>101</td>
<td>106.0</td>
<td>111</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>2</td>
<td>466</td>
<td>103</td>
<td>233.0</td>
<td>203</td>
</tr>
<tr>
<td>52</td>
<td>2</td>
<td>267</td>
<td>120</td>
<td>133.5</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>21351</td>
<td>45</td>
<td>205.3</td>
<td>512</td>
</tr>
</tbody>
</table>
SVY COMMAND IN STATA

```
.svymean SRATEHLT
Survey mean estimation

pweight: FWGTR  Number of obs =  21351
Strata: stratum  Number of strata =  52
PSU: secu  Number of PSUs =  104
Population size = 68027033

-------------------------------------------------------------
     Mean |  Estimate Std. Err.  [95% Conf. Interval]  Deff
-------------+--------------------------------------------------
    SRATEHLT |  2.839533   0.155449   2.808534   2.870532  3.670186
-------------------------------------------------------------
```

PROBLEMS WITH SUBSETTING

```
.xi:svyreg SRATEHLT i.RACE if MEXICAN
i.RACE (naturally coded; _RACE_0 omitted)
  note: _RACE_2 dropped due to collinearity
stratum with only one PSU detected
r(460);

.svyst if MEXICAN
pweight: FWGTR
Strata: stratum
PSU: secu

<table>
<thead>
<tr>
<th>Strata</th>
<th>$#PSUs</th>
<th>#Obs</th>
<th>min</th>
<th>mean</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>-----</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.0</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>54</td>
<td>23</td>
<td>27.0</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>46</td>
<td>19</td>
<td>23.0</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>12</td>
<td>6</td>
<td>6.0</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>3.5</td>
<td>5</td>
</tr>
</tbody>
</table>
```

HRS
AVOID THIS ERROR BY USING SUBSETTING COMMAND

STATA:

Svy, subpop (var)


SAS:

Use DOMAIN statement in survey procedure


SURVEY PROCEDURE IN SAS

```
proc surveyfreq data=hrs2004;
    cluster secu;
    strata stratum;
    weight jwgtr;
    table gender*arthritis04/col row chisq;
run;
```

The SURVEYFREQ Procedure

Data Summary

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Strata</td>
<td>56</td>
</tr>
<tr>
<td>Number of Clusters</td>
<td>112</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>20104</td>
</tr>
<tr>
<td>Number of Observations Used</td>
<td>18565</td>
</tr>
<tr>
<td>Number of Obs with Nonpositive Weights</td>
<td>1539</td>
</tr>
<tr>
<td>Sum of Weights</td>
<td>81582050</td>
</tr>
</tbody>
</table>
USING THE COMPLEX SAMPLE MODULE IN SPSS

Step 1: Set up a CS plan file prior to running analysis

CSPLAN ANALYSIS
/PLAN FILE=C:\MyFiles\HRS_workfilename.csaplan'
/PLANVARS ANALYSISWEIGHT=JWGTR
/SRSESTIMATOR TYPE=WR
/PRINT PLAN
/DESIGN STRATA=STRATUM CLUSTER=SECU
/ESTIMATOR TYPE=WR.

**Note that the “design” and “analysisweight” variables are listed in bold.

USING THE COMPLEX SAMPLE MODULE IN SPSS

Step 2: Run your procedure invoking the CS module and plan file

Complex Samples Descriptives.
CSDESCRIPTIVES
/PLAN FILE=C:\MyFiles\HRS_workfilename.csaplan'
/SUMMARY VARIABLES=FAMINC
/MEAN
/STATISTICS SE CI(95)
/MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
HOW IMPORTANT ARE THE WEIGHTS?

- Varies by variable and type of analysis
  - Crucial in descriptive analysis
    - Most important for variables closely related to the sample design (race and ethnicity) and any variables strongly correlated with these
  - Controversial in structural modeling
    - If model is properly specified, it may not reduce bias
    - It can introduce bias if weights are correlated with dependent variable
    - It can reduce efficiency by introducing variance

ADVICE ON USING WEIGHTS
CROSS-SECTIONAL ANALYSIS

- Respondent-level analysis: XWGTR
- Household-level analysis: XWGTHH
- Nursing home residents: XWGTNH
- Supplemental studies and enhanced face-to-face interview subsample weights that can found in the Tracker file include, for example:
  - XWGTR_DB (diabetes study)
  - XBIOWGTR (biomarkers)
  - XPMWGTR (physical measures)
  - XLBWGTR (psychosocial leave-behind)
  - Others can be found with their data (ADAMS, CAMS, etc.)

ADAMS = Aging, Demographics, and Memory Study
CAMS = Consumption and Activities Mail Survey

LONGITUDINAL ANALYSIS

- Longitudinal analyses:
  - Terminal year weight is appropriate for retrospective analysis where the goal is to model or describe the histories of individuals or households in the sample as of the terminal year
  - Base-year weight is more appropriate for prospective analyses where the goal is to model or describe the future experiences of the base-year population
WHY NON-POSITIVE (ZERO) WEIGHT?

Variable specifying reason for zero weight also provided for each wave

- Xwhy0wgt thru 2002
- Xwhy0hwt and Xwhy0rwt for 2004+

Why zero respondent weight?
0. Eligible, should have non-zero weight
1. Deceased
2. In a nursing home
3. Not cohort-eligible
4. Non-original sample member, not coupled with an original sample member
5. Noninterviews
6. Dropped by study
7. Not in sample this wave

Why zero household weight?
0. Eligible, should have non-zero weight
1. Only or both members deceased
2. Only or both members were in nursing homes
3. No cohort-eligible member in household
4. No original sample member in household
5. Only or both members were noninterviews
6. Dropped by study
7. Not in sample this wave

WHAT ABOUT COMBINING CORE AND SUPPLEMENTAL DATA?

- Think about your research question and the relevant population
- Choose the weight from the more restrictive population
  - E.g., respondents who got the physical measures (measured height and weight, timed walk, balance etc.) are a subset of core respondents, so use XPMWGTR
WHAT ABOUT MORE THAN ONE SUBSAMPLE?

- Example: Psychosocial leave-behind questionnaire and physical measures
- Choose the weight from the more restrictive population, or if same reference population, start with the weight from the larger sample size
- Carry out a weighting class adjustment or use propensity score weighting to adjust the selected weights for those cases in both samples

TWO HALF SAMPLES FROM CONSECUTIVE EFTF WAVES

- Half-sample weights sum to population size in each wave
  - This is only an issue if you want to report number of individuals in the population with some characteristic
  - Otherwise:
    - Rename your variables (drop the prefix)
    - Stack the data
    - More on this in the next webinar!
REFERENCES

- Documentation>Survey Design>Weight Information
  - Check out all four documents
- Heeringa, West, & Berglund. Applied Survey Data Analysis (2nd Ed)
  - Excellent general reference
- The Analysis Factor on complex sample design (theanalysisfactor.com)
  - For good plain explanations!
Questions?

- We will not be using the “raise hand” feature today
- Please use the “questions” feature accessible on the right side of your screen
- If we do not get to all of the questions today, we will email responses after the webinar

Thank You

- Other Health and Retirement Study videos on GSA’s YouTube
- Introduction to the Health and Retirement Study
- Data on Cognition
- Biomarkers and Physical Measures Data

hrsquestions@umich.edu

www.geron.org/webinar
Webinar Evaluation

In an effort for continual improvement, we would like to hear your thoughts. Please provide feedback by clicking the survey link at the end of the webinar.

Thank you again and we hope you enjoyed the program!

Thank You

GSA: Advancing Innovation in Aging
• The nation’s oldest and largest interdisciplinary organization devoted to research, education, and practice in the field of aging
• 5,500+ interdisciplinary members around the world touching all facets of aging
• Mission
  • Promote multi- and interdisciplinary research in aging
  • Translate and disseminate research findings
  • Promote/advocate for education/awareness on aging across disciplines
  • Foster application of research into policy development
• To view other GSA webinars, visit geron.org/webinar