The Survey of Income and Program Participation (SIPP)

* Critical Issues for Data Analysis using the SIPP

H. Luke Shaefer
University of Michigan School of Social Work
National Poverty Center

This presentation is part of the NSF-Census Research Network project of the Institute for Social Research at the University of Michigan. It is funded by National Science Foundation Grant No. SES 1131500.

Setting Up a SIPP Analysis

- The SIPP's complexity calls for extra care & obsessiveness
- Recommendation: Construct all analyses in a permanent set of STATA do files:
  1. A dataset construction file that loads in wave files, other data files, and drops unnecessary variables (leaving your core wave files unchanged)
  2. A variable construction file that reshapes variables (and maybe file format) as you need them for the analysis
  3. A set of analysis files that log and run each analysis (table 1; table 2; and so on)
- Using this structure makes it easier to:
  - Add variables to your dataset and reconstruct
  - Find mistakes—because you will know where to look
  - Re-run analyses and precisely replicate your results

SIPP Critical Issue: What's the Unit of Analysis?

- **Individuals:** Each individual sample member
- **Households:** "a group of persons who occupy a housing unit"
  - Includes: Families, a group of friends sharing a house, two unrelated families, co-housed, an unmarried mother and boyfriend
- **Family:** 2+ people related by birth, marriage, or adoption who reside together
  - See any potential problems here, given family complexity?
  - Easier to focus on dyads (mother/child) or a focal person
- **Related subfamily:** A nuclear family related to, but not including the household reference person
- **Unrelated subfamily:** A nuclear family that is not related to the household reference person
- **Note:** For all but the individual-level, you will have *multiple records* in a reference month for each member of the unit

Identifying Your Unit of Analysis

Unit of Analysis	Unique Identifier	Description
Individual (>= 1996)	ssuid + epppnum	sampling unit ID + person number
Individual (< 1996 panel)	suid + entry + pnum	sampling unit + entry address + person number
Household	ssuid + shhadid	sampling unit ID + current address ID
Family	ssuid + shhadid + fid	sampling unit ID + current address ID + family ID
Subfamily	ssuid + shhadid + rsid	Sampling unit ID + current address ID + family ID for related/ unrelated subfamilies

Good practice to add spanel to any identifier when stacking panels
NOTE: Family IDs do not stay constant across months, so you can't use the identifier to track a specific family from month-to-month

Unit of Analysis: What Observations do you Need?

- **Individuals:** Keep all respondent observations in your sample universe
- **Households:** Keep 1 observation per household
  - Household heads are the "owner or renter of note"
  - Can change from month-to-month
  - Use errp = 1 | 2, or
  - household head number, ehrefper = epppnum
  - Make sure characters match each other
- **Families:** Keep 1 observation per family
  - efrefper = epppnum
  - Same process for subfamilies (esfrfper)
- Household/family/subfamily variables are recorded in each sample member's observation, making life easier

Ordering Observations Chronologically

- A respondent's observations are ordered by:
  - WAVE (swave), then REFERENCE MONTH (srefmon)
  - Sort ssuid epppnum swave srefmon to order your dataset by unique respondent, then observations chronologically
- Note that in any given reference month, observations coming from 4 calendar months
- Can also order observations by calendar month and year
  - rhcalmn = Calendar month
  - rhcalyr = Calendar year
  - Note that in any given calendar month, observations are coming from 4 reference months

Creating a Year-Month Marker

- Syntax by Matt Rutledge
  - He uses Stata's time series functions now, but I still find this syntax useful

/* Reformat month and year variables to make one time-marking variable */

```
#delimit;

gen zero = 0;

egen tempmo = concat(zero rhcalmn);

tostring rhcalmn, generate(rhcalmn2);

replace tempmo = rhcalmn2 if rhcalmn > 9;

egen month = concat(rhcalyr tempmo);

drop tempmo rhcalmn zero rhcalmn2;
```

SIPP Critical Issue: Dealing with Seam Bias

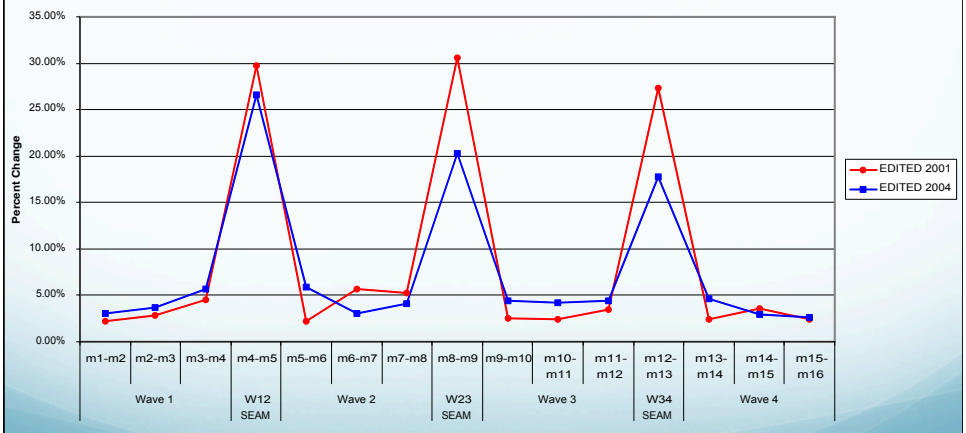
- Best known limitation of the SIPP is its "seam bias"
- Survey responses are most accurate in reporting months (month of the interview)
- Thus, a disproportionate number of transitions/changes occur between reference month 4 of wave *t*, and reference month 1 of wave *t+1*
  - Worse for some variables, better for others (employment spells)
- This affects the precision of estimates, especially of duration models
- **But a starting note:** The SIPP's relatively short seam could be considered a strength, rather than a weakness!
- Rotation groups mean in any calendar-year month you've got observations from all 4 rotation groups (on and off seam)

2004 Panel: Improved, but Still Visible, Seam Bias (Moore, 2008)

- With the 2004 panel, Census began to use dependent interviewing (DI) more comprehensively than before:
  - Prompting respondents with affirmative responses from the previous wave's reference month; and
  - Utilizing responses from the month in which the interview itself occurred
    - Current month responses were first collected in 1996 when Census transitioned to computer-assisted survey administration, but not yet utilized in the survey
  - DI reduced—but did not eliminate—seam bias
  - And this reduced variability in outcomes such as earnings/ incomes from wave-to-wave

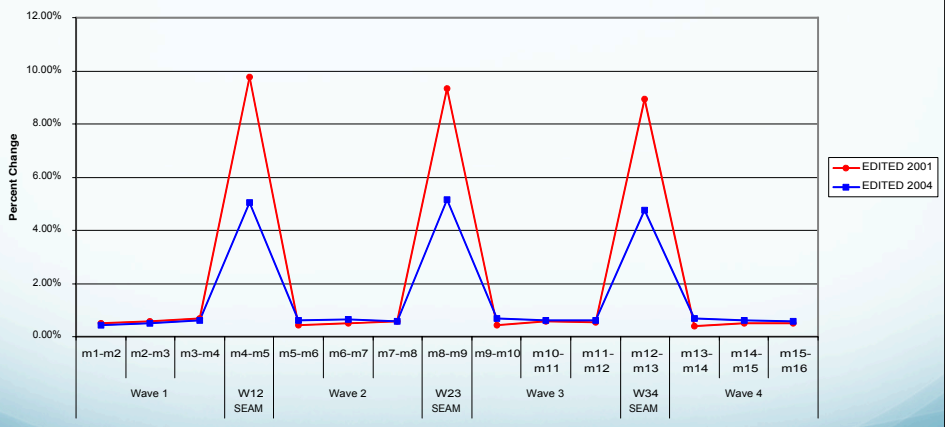
Improved, but Still Visible, Seam Bias (Moore, 2008)

Figure 1.A.4: MONTH-TO-MONTH AFDC/TANF CHANGES (unweighted) IN THE 2001 AND 2004 SIPP PANELS, WAVES 1 - 4
(edited data using cases with any AFDC/TANF receipt interviewed in all 4 waves)



Improved, but Still Visible, Seam Bias (Moore, 2008)

Figure 1.B.1: MONTH-TO-MONTH PRIVATE HEALTH INSURANCE COVERAGE CHANGES (unweighted) IN THE 2001 AND 2004 SIPP PANELS, WAVES 1 - 4
(edited data using cases with any private coverage interviewed in all 4 waves)



- Seam bias is much improved, but still visible
- Must be addressed in your research

Strategies for Dealing With Seam Bias

- **Option 1:** Add an indicator variable for reporting months in your models
  - Recommended by Ham, Li & Shore-Sheppard, 2007 as the safest practical method
- **Option 2:** Keep only reporting month observations
  - keep if srefmon == 4
  - Treat the data longitudinally as 4 month snap shots
  - If you hope to do exact durations in months, this will be imprecise
- **Option 3:** Collapse data into person-wave observations
  - Requires some arbitrary decisions when turning monthly data into four-month values
- **Option 4:** Predict mis-reporting and adjust accordingly (Ham et al. method), see technical paper

SIPP Critical Issue: Using appropriate Weights

- For representative estimates, weights are important because the SIPP:
  1. oversamples from high poverty areas (which is good!)
  2. is stratified, not purely random
- Some (usually economists) argue weights are unnecessary for multivariate estimates that control for the characteristics of oversampled populations
  - Census experts shudder uncontrollably when they hear this argument made…
- Often weights do not affect point estimates appreciably, but sometimes they do!!!
- Protect yourself: do it both ways
  - With/without weights
- The longer your recall period, the more important weights become because of attrition issues

Weights: Which to Use?

Unit of Analysis (Monthly Estimates)	Weight
Individual	wpfinwgt
Household	whfnwgt
Family	wffinwgt
Subfamily	wsfinwgt

- Or, take the person weight of the householder/family head, which will stay more stable over time
- Use of these weights adjusts point estimates but does not adjust standard errors (except if you use replicate weights)
- Presentation by Tracy Mattingly makes the case for using replicate weights and provide syntax to use them to adjust both point estimates and standard errors

Weighting for Longitudinal Analysis

- Attrition presents challenges when it comes to accurately modeling longitudinal outcomes
  - Less-advantaged respondents disproportionately drop from the sample over time due to residential instability
  - If you use the sample weight in t, but restrict to individuals in the sample in t+1, your weights may no longer be representative
- "Longitudinal" life is messy: (people die)
- One option for lag/lead variables is to use the monthly weight in the final month of your study period
  - So use t+1 weights rather than t
  - Then you are weighting on a cross-sectional sample, looking retrospectively
  - Even, still, you may experience problems with non-random entrance into the sample (probably minor)

Longitudinal Weights

- For longitudinal analyses for a calendar year, or the duration of the panel, use longitudinal weights
- These track sample members who remain "in universe" for the duration of the time period
- These weights adjust for attrition by increasing weights on sample members representing sub-populations who attrit (a word?)
  - But this means that sample cells for small subpopulations can get VERY small
- Merge into core using unique individual ID (ssuid + epppnum)
- Convert monthly responses into year/panel data using unique identifiers

  keep if rhcalyr == 2009

  bysort ssuid epppnum: egen annearnings = total(tpearn)

SIPP Critical Issue: Imputation

- When a respondent refuses or is unable to answer a question, Census will impute a value for them
  - Oversimplified description: Census uses values from other, similar respondents
- **Upside:** The SIPP public use data files have little missing data
- **Downside:** We sometimes question the accuracy of imputed data
- (Generally) rising rates of data imputation are a concern for the accuracy of household survey data

Imputation

- Ways of dealing with imputation:
  1. Use only non-imputed data
     - This creates numerous problems and is not a practice that Census endorses
     - My recommendation is to do this as a sensitivity test at most
     - Difficult to do with some measures recoded from a series of variables
  2. If using 2+ panels, compare differences between the **end** of 1 panel and the **beginning** of the next (maybe wave 2)
     - Imputation is generally LOWEST at the beginning of the panel and HIGHEST at the end
  3. Alternative imputation: You can re-impute using multiple imputation or another technique

SIPP Critical Issue: Adjusting your Standard Errors

- The SIPP's stratified sample design leads to overly narrow standard errors
- Can lead to misleading labeling of statistical significance
- This must be accounted for in your analysis. Choices for doing so that have precedence in the literature:
  1. Using replicate weights (see Tracy Mattingly's lecture)
  2. Using STATA's svyset function
  3. Robust clustering of standard errors by state
  4. Generating bootstrapped standard errors
     - no good way to do this with weights
     - Not an approach endorsed by Census

Adjusting your Standard Errors

OPTION 2: USE STATA'S SVYSET TO ADJUST FOR COMPLEX SURVEY DATA

Example: Predicting Earnings by Education Level using 2008 panel, wave 1

(Oversimplified, silly example)

```
keep if tage > 17 & tage < 65

svyset ghlfsam [pw = wpfinwgt], strata(gvarstr)

svy: reg tpearn i.eeducate
```

Point estimate associated with a master's degree relative to less than a 1st grade education:

$8,129 (350.95)

Adjusting your Standard Errors

OPTION 1: ROBUST CLUSTERING OF STANDARD ERRORS BY STATE

Example: Predicting Earnings by Education Level using 2008 panel, wave 1

(Oversimplified, silly example)

```
Keep if tage > 17 & tage < 65

reg tpearn i.eeducate [pw = wpfinwgt], vce
(cluster  tfipsst)
```

Point estimate/se associated with a master's degree relative to less than a 1st grade education (monthly income):

$8,129 (367.92)

Adjusting your Standard Errors

OPTION 3: USING BOOTSTRAPPING WITH REPLACEMENT

Example: Predicting Earnings by Education Level using 2008 panel, wave 1

(Oversimplified, silly example)

- **Note:** No good way I know of in Stata to bootstrap in the context of a complex stratified sample design

```
keep if tage > 17 & tage < 65

bootstrap, reps(500): reg tpearn i.eeducate
```

Point estimate associated with a master's degree relative to less than a 1st grade education:

$8,348 (168.24)