2017 SURVEY AND DIARY OF CONSUMER PAYMENT CHOICE

Sampling and Weighting

(Marco Angrisani, USC, 1/31/2018)

1. UAS Sample Description

The UAS is a nationally representative panel of U.S. households recruited through Address Based Sampling (ABS). Eligible individuals are all adults in the contacted household aged 18 and older. Sampling in the UAS is done in batches. The first batch (batch 1) is a simple random sample of individuals from the ASDE Survey Sampler database. Subsequent recruitment batches (batches 5-12) are selected based on an algorithm developed by Center for Economic and Social Research (CESR) researchers called Sequential Importance Sampling (SIS). This is a type of adaptive sampling that allows to refresh the panel in such a way that its demographic composition moves closer to the population composition.

Specifically, before sampling an additional batch, the SIS algorithm computes the unweighted distributions of specific demographic characteristics (e.g., sex, age, marital status and education) in the UAS at that point in time. It then assigns to each zip code a non-zero probability of being drawn, which is an increasing function of the degree of "desirability" of the zip code. The degree of desirability is a measure of how much, given its population characteristics, a zip code is expected to move the current distributions of demographics in the UAS towards those of the U.S. population. The implementation of the SIS algorithm implies that the marginal probability of drawing each zip code depends on the composition of the UAS panel at a particular point in time, but also on the unknown response probabilities of selected households in that zip code. Hence, the marginal probability of drawing each zip code is not known ex ante and cannot be used to construct design weights. The UAS weighting procedure features base weights to correct for the unequal sampling probabilities generated by the SIS algorithm.

The UAS also includes three special purpose samples – a sub-panel of Native Americans, a sub-panel of Los Angeles County residents and a sub-sample of California residents – for which different sampling procedures are adopted. The sample of Native Americans (batches 2 and 3) is recruited through ABS, targeting zip codes with a higher proportion of Native Americans. In this case, eligible individuals are all Native American adults in the contacted household, aged 18 and older. Recruitment of the first special purpose sample of Los Angeles County residents (batch 4) is based on birth records information from the State of California. Later special purpose samples of Los Angeles County residents (batches 13 and 14) are again recruited through ABS. The special purpose sample of California residents is recruited through ABS.

2. SCPC and DCPC Sample Selection

For the 2017 Survey of Consumer Payment Choice (SCPC) and Diary of Consumer Payment Choice (DCPC), only UAS members from nationally representative batches were invited to take part in the study. The selection procedure was carried out in two steps. In the first step, panel members were asked about their willingness to participate in a two-phase study consisting of the SCPC and the DCPC. In the second step, those who consented were invited to take the SCPC first and then the DCPC at designated dates. The SCPC was fielded on September 19, 2017. The fielding period for the DCPC was defined accordingly to run from September 28, 2017 to November 2, 2017.

The number of UAS members available at the time of the sample selection (August 2017) who were part of the Nationally Representative core sample was 4,759. These respondents were assigned to two groups depending on whether or not they had previously participated in the study. The first group of former participants had 3,677 respondents and was first invited to take the consent survey. The second group of UAS members who had not participated in the study before had 1,082 respondents and was invited to take the consent survey two weeks later. The consent survey was completed by 3,293 respondents, of which 3,158 were willing to participate in both the SCPC and the DCPC, and 135 were not willing to participate in the study.

Out of the 3,158 who were invited to take the SCPC, 3,099 completed the survey for a response rate of 98%. Out of the 3,099 who completed the SCPC, 2,871 participated in the DCPC, but 36 only completed "day 0" of the diary. Excluding the latter, the response rate is about 91%.

3. Weighting Procedure

Sample weights for typical UAS surveys are constructed in two steps. In a first step, a *base weight* is created to account for unequal probabilities of sampling zip codes produced by the SIS algorithm and to reflect the probability of a household being sampled, conditional on its zip code being sampled. In a second step, *final post-stratification weights* are generated to correct for differential non-response rates and to bring the final survey sample in line with the reference population as far as the distribution of key variables of interest is concerned.

3.1. **Categorization and imputation of variables**

As far as the UAS sample is concerned, we use demographic information taken from the most recent “My Household” survey, which is answered by the respondent every quarter. With the exception of age and number of household members, all other socio-demographic variables in the “My Household” survey are categorical and some, such as education and income, take values in a relatively large set. We recode all the variables used in the weighting procedure into new categorical variables with no more than 5 categories. The aim of limiting the categories is to prevent these variables from forming strata containing a very small fraction of the sample (less than 4-5%), which may cause sample weights to exhibit considerable variability. The categorization of variables used for the weighting procedure follows the same definitions adopted for the 2014-2016 SCPC/DCPC, in order to ensure comparability across years. The list of recoded categorical variables used in the weighting procedure is reported in Table 1.

Table 1: List of Recoded Categorical Variables Used within the Weighting Procedure

Recoded Variable	**Categories**
gender	1. Male; 2. Female
age_cat *age_cat2*	1. 18-34; 2. 35-44; 3. 45-54; 4. 55-64; 5. 65+ 1. 18-44; 2. 45-64; 3. 65+
bornus	0. No; 1. Yes
citizenus	0. No; 1. Yes
marital_cat	1. Married; 2. Separated/Divorced/Widowed; 3. Never Married
education_cat	1. High School or Less; 2. Some College/Assoc. Degree; 3. Bachelor or More
hisplatino	0. No; 1. Yes
race_cat	1. White; 2. Non-White
work_cat	1. Working; 2. Unemployed; 3. Retired; 4. On leave, Disabled, Other
hhmembers_cat	1. One Member; 2. Two Members; 3. Three or More Members

hhincome_cat	1. <$30,000; 2. $30,000-$59,999; 3. $60,000-$99,999; 4. $100,000+
hhincome_cat2	1. <$35,000; 2. $35,000-74,999; 3. $75,000+

Before implementing the weighting procedure, we employ the following imputation scheme to replace missing values of recoded socio-demographic variables.

- We do not impute gender. Hence, respondents with missing gender are not assigned a sample weight. No respondent in the 2017 SCPC and DCPC samples has missing gender.
- When actual age is missing, the variable *agerange*, available in the “My Household” survey, is used to impute *age_cat*. If *agerange* is also missing, the variable *age_cat* is assigned the mode for males or females, depending on the respondent’s gender.
- For binary indicators, such as *bornus*, *citizenus*, and *hisplatino*, missing values are imputed using a logistic regression.
- For ordered categorical variables, such as *education_cat*, *hhmembers_cat*, *hhincome_cat* and *hhincome_cat2*, missing values are imputed using an ordered logistic regression.
- For non-ordered categorical variables, such as *marital_cat*, *race_cat* and *work_cat*, missing values are imputed using a multinomial logistic regression.

Imputations are performed sequentially. That is, once *age_cat* has been imputed (if missing), the variable with the smallest number of missing values is the first one to be imputed by means of a regression featuring *gender* and *age_cat* as regressors. This newly imputed variable is then added to the set of regressors to impute the variable with the second smallest number of missing values. The procedure continues in this fashion until the variable with the most missing values (typically household income) is imputed using information on all other socio-demographic variables.

The final 2017 SCPC and DCPC data sets contain a binary variable, *imputation_flag*, indicating whether any of the recoded socio-economic variables listed in Table 1 has been imputed.

2.2. **Post-stratification Weights**

The execution of the sampling process for a survey is typically less than perfect. Even if the sample of panel members invited to take a survey is representative of the population along a series of dimensions, the sample of actual respondents may exhibit discrepancies because of differences in response rates across groups and/or other issues related to the fielding time and content of the survey. Weighting is therefore needed to align the final survey sample to the reference population as far as the distribution of key variables is concerned. We perform **iterative marginal weighting** and assign survey respondents weights such that the weighted distributions of specific socio-demographic variables in the survey sample match their population counterparts (benchmark or target distributions).

The benchmark distributions against which the 2017 SCPC and DCPC are weighted are derived from the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) administered in March of 2017. The reference population is the U.S. population of those aged 18 and older, excluding institutionalized individuals and military personnel.

We adopt a **raking algorithm** to generate post-stratification weights. This procedure involves the comparison of target population relative frequencies and actually achieved sample relative frequencies on a number of socio-demographic variables independently and sequentially. More precisely, starting from an initial weight of one, at each iteration of the algorithm weights are proportionally adjusted so that the distance between survey and population marginal distributions of each selected socio-demographic variable (or raking factor) decreases. The algorithm stops when survey and population distributions are perfectly aligned. A maximum of 50 iterations is allowed for perfect alignment of survey and population distributions to be achieved. If the process does not converge within 50 iterations, no sample weights are returned and attempts using different raking factors are made.

2.3. **Trimming**

Our raking algorithm trims extreme weights in order to limit variability and improve efficiency of estimators. We follow the general weight trimming and redistribution procedure described by Valliant, Dever and Kreuter (2013). Specifically, indicating with $w_{i,raking}$ the raking weight for respondent *i* and with $\bar{w}_{raking} = \frac{1}{N}\sum_{i=1}^{N} w_{i,raking}$ the sample average of raking weights,

I. We set the lower and upper bounds on weights equal to $L = 0.25\bar{w}_{raking}$ and $U = 4\bar{w}_{raking}$, respectively. While these values are arbitrary, they are in line with those described in the literature and followed by other surveys (Izrael, Battaglia and Frankel, 2009).

II. We reset any weights smaller than the lower bound to L and any weights greater than the upper bound to U:

$$w_{i,trim} = \begin{cases} L & w_{i,raking} \leq L \\ w_{i,raking} & L < w_{i,raking} < U \\ U & w_{i,raking} \geq U \end{cases}$$

III. We compute the amount of weight lost by trimming as $w_{lost} = \sum_{i=1}^{N} w_{i,raking} - w_{i,trim}$ and distribute it evenly among the respondents whose weights are not trimmed.

While raking weights can match population distributions of selected variables, trimmed weights typically do not. We therefore iterate the raking algorithm and the trimming procedure until a set of post-stratification weights is obtained that respect the weight bounds and align sample and population distributions of selected variables. This procedure stops after 50 iterations if an exact alignment respecting the weight bounds cannot be achieved. In this case, the trimmed weights will ensure the exact match between survey and population relative frequencies, but may take values outside the interval defined by the pre-specified lower and upper bounds.

2.4. **Final Post-stratification Weights**

Indicate with $w_{i,post}$ the post-stratification weight for respondent *i*, obtained after iterating the raking algorithm and the trimming procedure as described above

The final 2017 SCPC and DCPC post-stratification weights are expressed relative to their sample mean. That is:

$$relw_{i,post} = \frac{w_{i,post}}{\left(\frac{1}{N} \sum_{i=1}^{N} w_{i,post}\right)},$$

where *N* is the survey sample size.
These relative post-stratification weights average to 1 and sum to the survey sample size *N*.

One respondent (uasid=141000007) receives a weight of 0 by Boston Fed’s request. Four other respondents receive a weight of 0 because their base weight is 0.

3. Produced Sample Weights

We produce general weights for the SCPC and general, day-of-the-week and daily weights for the DCPC. General weights in both 2017 SCPC and DCPC and day-of-the-week weights in the DCPC are generated using the following set of raking factors:

- *gender x race_cat*
- *gender x age_cat*
- *gender x education_cat*
- *hhmembers_cat x hhincome_cat*

The same set of raking factors was adopted to produce general sample weights for the 2014-2016 SCPC/DCPC. Under this specification, both the raking and the trimming algorithms converge within the maximum number of allowed (50) iterations.

Because of the limited number of respondents taking the diary at specific days, daily weights for the DCPC are generated using a reduced set of raking factors, namely:

- *gender x age_cat2*
- *education_cat*
- *hhincome_cat2*

Again, this set of variables is the same as the one used for the 2014-2015 DCPC daily weights so to ensure comparability. Under this specification, the raking algorithm converges within the maximum number of allowed (50) iterations. We do not apply trimming to daily weights.

The complete list of weights and auxiliary variables provided with the final 2017 SCPC and DCPC data sets is reported below.

2017 SCPC:

- *imputation_flag*

  A binary variable indicating whether any of the variables listed in Table 1 has been imputed.
- *base_weight*

  Base weight.
- *final_weight*

  Final post-stratification weight.

2017 DCPC:

(note: the DCPC data set is in "long form" with 4 diary days (day 0-3) for each respondent)

- *day_week*

  Variable indicating the day of the week:

  0 = Sunday

  1 = Monday

  2 = Tuesday

  3 = Wednesday

  4 = Thursday

  5 = Friday

  6 = Saturday
- *imputation_flag*

  A binary variable indicating whether any of the variables listed in Table 1 has been imputed.
- *base_weight*

  Base weight.
- *final_weight*

  Final post-stratification weight for every diarist.
- *final_weight_dow*

  Final day-of-the-week weight (within the month of October).
- *final_weight_day*

  Final daily weight.