UnderStandingAmericaStudy

WEIGHTING PROCEDURE FOR UAS COVID SURVEYS

USC Dornsife Center for Economic and Social Research

UNDERSTANDING AMERICA STUDY (UAS) COVID SURVEYS

WEIGHTING PROCEDURE

May 2020

Contents

INTRODUCTION

This document provides a brief description of the weighting procedure used to create sample weights for the Understanding America Study Coronavirus in America ("COVID") surveys. The weighting procedure described in this document follows closely the general weighting procedure adopted for UAS surveys. Please refer to this page for more details.

1. SAMPLING

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 carried out in batches (more details are available here). There are currently 20 batches, targeting either the U.S. population at large, or specific subsets of it, such as the population of Native Americans, California residents, and Los Angeles County residents.

All UAS batches, except batch 4 (a simple random sample from a list of women who had given birth in Los Angeles County between 2009 and 2012), were invited to take COVID surveys.

UAS recruitment batch 1 was a simple random sample of the U.S. adult population. All subsequent batches use a two-stage sample design, in which zip codes are drawn first, and then households are randomly drawn from the sampled zip codes. This two-stage sample design is based on an adaptive sampling algorithm and allows to refresh the panel in such a way that its demographic composition moves closer to the reference population's composition. The implementation of this algorithm implies unequal sampling probabilities across individuals, depending on the demographic characteristics of their zip code of residence and the demographic composition of the UAS panel at the time the recruitment of a new batch is carried out. The weighting procedure corrects for these unequal sampling probabilities.

2. WEIGHTING

Sample weights for all UAS COVID surveys are wave-specific. They are provided with each UAS survey, whether a National or a Los Angeles County survey data set, and are meant to make each survey data set representative of the reference population with respect to a pre-defined set of socio-demographic variables. Sample weights are constructed in two steps. In a first step, a *base weight* is created to account for unequal probabilities of sampling UAS members generated by the adaptive sampling algorithm. In a second step, *final poststratification 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.

2.1. Step 1: Base Weights

When computing base weights, the unit of analysis is a zip code. We estimate a logit model for the probability that a zip code is sampled as a function of its characteristics, namely Census region, urbanicity, population size, as well as sex, race, age, marital status and education composition. Estimation is carried out on an American Community Survey (ACS) file that contains 5-year average characteristics at the zip code level.[1] The outcome of this logit model is an estimate of the marginal probability of a zip code being sampled, which, given the implementation of the adaptive sampling algorithm described above, is not known *ex ante*.

We indicate by π_k the logit estimated probability of sampling zip code k. The probability of sampling household h after drawing zip code k is the ratio of the number of households sampled divided by the number of households in the zip code. We indicate this by $\pi_{h|k}$. Hence, the marginal probability that household h from zip code k is sampled is $\pi_{hk} = \pi_{h|k} \times \pi_k$.

The base weight is a zip code level weight defined as:

$$w_{hk}^{base} = \Lambda \times \frac{1}{\pi_{hk}}$$

[1] Strictly speaking, all files from the U.S. Census Bureau use "zip code tabulation area" (zcta), which is based on, but not identical to, USPS's definition of zip codes. We ignore the distinction between the two.

where the constant Λ is chosen such that the sum of the base weights is equal to the number of sampled households. A comprehensive discussion of how base weights are computed is available here.

UAS members are assigned a base weight, computed as described above, depending on the zip code where they reside at the time of recruitment.

2.2. Step 2: Poststratification 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. A second layer of weighting is therefore needed to align the final survey sample to the reference population as far as the distribution of key variables is concerned.

In this second step, we perform **raking weighting** (also known as iterative marginal weighting), starting from the base weights, w_{hk}^{base}, described in the previous section. With this, we assign poststratification weights to survey respondents 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 UAS COVID surveys are weighted are derived from the Basic Monthly Current Population Survey (CPS).[2] We use the 6 most recent available monthly CPS at the time a UAS COVID survey is completed. This ensures a minimum gap between the period of survey completion and the period benchmark distributions refer to.

The reference population for UAS COVID National surveys is the U.S. population of adults, age 18 or older, excluding institutionalized individuals and military personnel. The reference population for UAS COVID Los Angeles County surveys is the Los Angeles County population of adults, age 18 or older, excluding institutionalized individuals and military personnel.

[2] We rely on IPUMS-CPS, University of Minnesota, https://ipums.org/.

2.3. Categorization and Imputation of Variables

For poststratification weighting purposes, we use demographic information taken from the most recent *My Household* survey, which is answered by all active UAS members every quarter. All socio-demographic variables in the *My Household* survey are categorical, but some, such as age and education take values in a relatively large set. We recode all socio-demographic variables considered for poststratification into new categorical variables with no more than 5 categories. The aim of limiting the number of categories is to prevent these variables from forming strata containing a very small fraction of the sample (less than 5%), which may cause sample weights to exhibit considerable variability. Recoded categorical variables used for poststratification and their corresponding definitions are provided in Table 1.

Table 1: Categorical Variables for Poststratification

Recoded Variable	Categories
Gender	1. Male; 2. Female
Age	1. 18-39; 2. 40-49; 3. 50-59; 4. 60+
Education	1. High School or Less; 2. Some College; 3. Bachelor or More
Race-ethnicity	1. White; 2. Black; 3. Others; 4. Hispanic; 5. Native American
*Census region (aug)**	1. Northeast; 2. Midwest; 3. South; 4. West, excluding California; 5. California, excluding Los Angeles County; 6. Los Angeles County

* These are obtained from the respondent's zip code of residence.

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

- Gender is obtained from administrative records.
- When age is missing, the age range available in the My Household survey is used to impute age categories. If the age range is also missing, age categories are imputed using gender-specific sample mode.
- Once age categories have been imputed (if missing), the variable with the fewest missing values is the first one to be imputed by means of a regression featuring gender and the age categories 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 is imputed using information on all other available socio-demographic variables.

For education and race-ethnicity, missing values are imputed using ordered logistic and multinomial logistic regression, respectively. Although UAS members report their current state of residence the *My Household* survey, we rely on administrative records of respondents' zip codes of residence to construct the variable Census region. This is treated as imputed whenever current state of residence is missing in the *My Household* survey.

Each UAS COVID survey data set includes a binary variable (imputation flag) indicating whether any of the recoded socio-economic variables used for poststratification has been imputed.

2.4. Raking/Trimming Algorithm

We adopt a raking algorithm to generate poststratification 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 the base weights, 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.

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).[3]

[3] Valliant, R., Dever, J. A., and Kreuter F., (2013) *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York.

Specifically, we define N the total sample size. Indicating with w_i^{rak} the raking weight for respondent $i = 1, \dots, N$, and with $\bar{w}^{rak} = \frac{1}{N}\sum_{i=1}^{N} w_i^{rak}$ the sample average of raked weights,

I. We set the lower (L) and upper (U) bounds on weights equal to the 10th and 90th percentile of the w_i^{rak} distribution, respectively. While there is no consensus on which threshold should be used for trimming, these are among those often mentioned in the literature and adopted by other surveys (Battaglia et al., 2009).[4]

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

$$w_i^{trim} = \begin{cases} L & w_i^{rak} \leq L \\ w_i^{rak} & L < w_i^{rak} < U \\ U & w_i^{rak} \geq U \end{cases}$$

III. We compute the amount of weight lost by trimming as $w^{lost} = \sum_{i=1}^{N_c}(w_i^{rak} - w_i^{trim})$ and distribute it equally among the respondents whose weights are not trimmed.

IV. If these new weights are all within the interval $[L, U]$, no further adjustment is performed. If any of these new weights are outside the interval $[L, U]$, the trimming procedure is repeated iteratively until all weights are within the interval $[L, U]$ or until the maximum number of 50 iterations is reached.

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 post-stratification weights are 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 raked weights will ensure an exact match of (weighted) survey relative frequencies to their population counterparts, but the weights will not be within the pre-determined bounds.

[4] Battaglia, M. P, Izrael, D., Hoaglin, D. C., and Frankel M. R., (2009) "Practical Considerations in Raking Survey Data." *Survey Practice*, 2009 (June). http://surveypractice.org/2009/06/29/raking-survey-data/.

2.5. Final Poststratification Weights

Indicate by w_i^{post} the final poststratification weight for respondent i, obtained by applying the raking algorithm to the base weights and after iterating the raking/trimming procedure as described above. Each weighted UAS COVID survey data set includes final poststratification weights relative to their sample mean. Formally, this is

$$w_i^{final} = \frac{w_i^{post}}{\left(\frac{1}{N}\sum_{j=1}^{N} w_j^{post}\right)}$$

Hence, relative final poststratification weights sum to the size of the sample of each UAS COVID survey and average to 1 within that sample.

3. WEIGHTING OUTPUT

The UAS is a nationally For the National UAS COVID surveys, we separate the sample into three groups based on geography. These are: 1) USA excluding California; 2) California excluding Los Angeles County; 3) Los Angeles County. We then implement the raking/trimming procedure described above separately for these three groups. The resulting poststratification weights for each of these sub-samples are finally scaled up to population size by multiplying them by the factor Pop_s/N_S, where S={USA excluding California, California excluding Los Angeles County, Los Angeles County}, and Pop_s and N_S are the population and sample size of each geographic sub-sample, respectively.

Each weighted UAS COVID survey data set includes the following variables:

- *base_weight*

  Base weights correcting for unequal sampling probabilities.

- *imputation_flag*

  A binary variable indicating whether any of the variables used for poststratification has been imputed.

- *cps*

  A variable indicating the CPS monthly surveys used to obtain benchmark distributions for poststratification.

- *final_weight*

  Relative final poststratification weights ensuring representativeness of the survey sample with respect to the following demographic variables: gender, race/ethnicity, age, education, and location.

  For National UAS COVID surveys, *final_weight* allows to match the distributions of the aforementioned demographic variables in the entire United States adult population. If one of the three specific geographic sub-samples is selected – USA excluding California, California excluding Los Angeles County, Los Angeles County – *final_weight* allows to match the distributions of the aforementioned demographic variables in each of these geographies.

  For Los Angeles County UAS COVID surveys, *final_weight* allows to match the distributions of the aforementioned demographic variables in the Los Angeles County population.

  Weighted statistics of interest can be produced using *final_weight* (the adjustment for unequal sampling probabilities provided by *base_weight* is incorporated in *final_weight*).