Rubrics for the Documentation of Sampling designs and Their Specification in Statistical Software Packages

Stas Kolenikov[1], Brady T. West[2], Peter Lugtig[3]

[1] Abt Associates, 6130 Executive Blvd, Rockville, MD 20852
[2] Survey Research Center, University of Michigan, 4118 ISR Building,
426 Thompson Street, Ann Arbor, MI 48109
3 Department of Methodology and Statistics, Utrecht University
Padualaan 14, 3584 CH Utrecht, the Netherlands

Abstract

We document our understanding of, and recommendations for, appropriate best practices in specifying the complex sampling design settings in statistical software that enables design-based analyses of survey data. We discuss features of complex sample survey data such as stratification, clustering, unequal probabilities of selection, and calibration, and outline their impact on estimation procedures. We demonstrate how statistical software treats them, and how the survey data providers can make data users' lives easier by clearly documenting accurate and efficient ways to make sure that their software properly accounts for the complex sampling design features. We provide rubrics that will aid complex sample survey data providers in aligning their level of documentation with best practices, and show how existing surveys and their documentation score based on these rubrics.

Keywords: design-based inference, population surveys, statistical software, complex samples, Total Survey Error

1. Principal sampling design features

A principal objective in survey research is to develop survey designs that minimize Total Survey Error (TSE; Groves and Lyberg 2010). Sampling and adjustment errors are two of the errors within the larger TSE framework that can be internally quantified in statistical software. When coverage and nonresponse errors can be estimated as well, there are possibilities to adjust errors in order to ensure that the analysis of the survey represents the larger population. If this is done well, the results from the survey analysis are asymptotically unbiased with respect to the sampling design, while uncertainty due to the various errors can be estimated as well. In this paper, we focus on the "big four" features of complex sampling designs: stratification, cluster sampling, unequal probabilities of selection, and weight adjustments. Each design feature is described in more detail below. Although we discuss the possible reasons why one would use a particular survey design, we refer to Groves et al (2011), Lohr (2010), Biemer & Lyberg (2003), Groves (2004) or other textbooks on survey design for a broader context and overview of sampling design decisions.

1.1 Stratification

Stratification divides the population and sampling frames into mutually exclusive groups (strata) before sampling. Common examples of strata include:

- Geographic regions for in-person samples;
- Diagnostic groups for patient list samples; and

- Industry, employment size and/or geographical regions for establishment samples.

Complex sampling designs employ stratification to:

- Oversample subpopulations of interest (e.g. ethnic minorities) if they can be identified on the frame(s);
- Oversample areas of higher concentration of the target rare population;
- Ensure specific accuracy targets in subpopulations of interest;
- Utilize different sampling designs/frames in different strata;
- Avoid outlying samples and spread the sample across the whole population;
- Optimize costs vs. precision via Neyman-Chuprow or more complicated allocations.

1.2 Cluster Sampling

Cluster, or multistage, sampling design consists of sampling groups of observation units (clusters), rather than the ultimate observation units directly. From a statistical efficiency viewpoint, this is a less desirable feature as clustering of units that have similar characteristics reduces precision of survey estimates. Common examples of randomly sampled clusters include:

- Geographic units (e.g., census tracts, enumeration districts) in face-to-face surveys;
- Entities in natural hierarchies (e.g. health care providers within practices within hospitals, or students within classes within schools).

Why do complex sampling designs employ cluster sampling?

- Complete lists of all units are not available, but survey statisticians can work with lists of administrative units (e.g., states, counties, Census tracts, enumeration districts) for which membership of the next stage sampling units can be clearly established;
- Reduce interviewer travel time/cost in face-to-face surveys;
- Substantive researchers have an analytic interest in multilevel modeling of hierarchical structures

1.3 Unequal Probabilities of Selection

In practice, sampling designs introduce unequal probabilities of selection for different sampling elements. From a solely statistical perspective, this is a less desirable feature as larger variances in weights across cases reduce the precision of survey estimates.

Complex sampling designs can assign unequal probabilities of selection to different population units to achieve a number of goals. Commonly, unequal probabilities result from implementation of a primary sample size target. First, when (smaller) subpopulations of interest (e.g., ethnic/racial minorities) that would not have sufficient sample sizes in an equal probability of selection method (epsem) sample are oversampled directly from lists, or indirectly oversampled by selecting geographic areas with a higher concentration of the target rare population, unequal probabilities of selection would result. Second, most samples for face-to-face surveys are designed with probability proportional to size (PPS) sampling at the first stages, with fixed sample size at the final stage to achieve an approximately epsem design. In many cases, however, the sample size at the final of selection (e.g. the size of a household) is unknown in advance, leading to different weights for the units of observation. Third, unequal probabilities are nearly inevitable in multiple frame sampling, where units can be sampled through several possible channels. In phone surveys, dual phone users, i.e., those who have both landline and cell phone service, are more likely to be selected than those who have cell only or landline only service.

1.4 Weight Adjustments

After the data are collected, survey statisticians further adjust the weights to make appropriate corrections (see Valliant and Dever, 2018, for details). These adjustments generally account for:

- Eligibility;
- Frame noncoverage;
- Frame overlap in multiple frame surveys;
- Statistical efficiency;
- Unit nonresponse.

1.5 Sampling is about doing the best job for the money!

At the end of the day, all of the complex sampling features described above are employed to save money and collect the data in more efficient ways. These efficiencies come with statistical tradeoffs, however. While the use of cluster samples would allow survey designers to save on travel costs, precision of the estimates will be worsened due to intracluster correlations. However, if travel costs are reduced by a factor of five, and the reduction in statistical efficiency is by a factor of two, then undoubtedly a cluster sampling design is the more economical one in units of precision per dollar. In most general population samples (except some European countries with excellent population registers), there is no access to the full population listing, forcing survey designers either to use area samples to gradually gain access to individuals, or use infrastructure created for a different purpose (phone communication or postal service) to contact potential respondents. Obtaining a full population list to sample from would be a prohibitively expensive exercise.

When studying populations that are subsets of the general population (e.g., families with children; religious minorities; military veterans; and many other special populations), survey statisticians may have multiple ways to reach these populations by screening out a larger, general population sample, or through the social systems associated with that population (e.g., daycare centers and schools to reach children). Those different frames may have different costs of identifying eligible units, but may have to be used in conjunction to ensure complete coverage of population (e.g. home schooled children can only be found in the general population sample that can be more expensive than a school-based sample) and correct inference. In studies of rare populations, the variance in weight factors will inevitably arise as a function of different screening rates, different coverage of the various frames used, and stratification of the frames oversampling areas of higher concentration of the population of interest that would allow to collect data less expensively. Finally, some adjustments are unavoidable to the extent that the real-world data collection challenges they correct for are unavoidable – such as unit nonresponse and nonresponse weight adjustments.

As a result of all the considerations above, population surveys employ complex sampling designs in their fieldwork. Data resulting from such complex surveys cannot be naively analyzed as is, and survey weights and possibly other elements of the complex sampling design have to be accounted for. Survey statisticians routinely compute weights for data users. These weights often take the form of a design weight that corrects for eligibility, frame overlap, and unequal selection probabilities in sampling. A separate nonresponse weight corrects for nonresponse, and sometimes for noncoverage errors in the frame used. In some surveys additional weights are provided for the purpose of doing cross-national comparisons (multi-country surveys) or longitudinal analysis (cohort or panel studies). For more information on how modern surveys are efficiently designed, and weights are computed, we refer the reader to Kalton, Flores-Cervantes (2003), Lohr (2010), Bethlehem (2010), Valliant, Dever, Kreuter (2013), Valliant and Dever (2018) or Kolenikov (2016). The weights included in a survey dataset should be accompanied with detailed documentation on how the weights were computed and how they should be used in practice by applied researchers. We have often found that the documentation of survey weights is inadequate. Sometimes, details on how the weights were designed are missing. More often, the description of the weights is sparse or very technical. This then leads to users not using weights at all, or using them incorrectly. West, Sakshaug and Aurelien (2016) have shown for example that analytic errors are prevalent in 145 analyses of the survey 'Scientists and Engineers Statistical Data System' (SESTAT). They reported that "… only 55% of the products incorporated the publicly-available sampling weights into the analyses, only 8% of the products accounted for the complex sampling features when estimating variances, and only 11% of the products presenting design-based analyses performed appropriate subpopulation analyses accounting for the complex sampling". In the medical domain, Khera et al (2017) reported that "a total of 79 [out of 120] sampled studies (68.3% [95%CI, 59.3%-77.3%]) among the NIS studies screened for eligibility did not account for the effects of sampling error, clustering, and stratification".

Ignoring survey design weights will lead to wrong inferences. Data users therefore will need to know why and how to use weights that are being provided with the public-use files of large survey data. Simultaneously, survey designers and methodologists need to document how these weights are being produced and provide guidance to users on how to use weights in practice. This paper therefore seeks to provide rubrics for how survey weights and sampling design settings should be documented for the ultimate survey data users. We will define a set of rubrics consisting of five main and two bonus elements, and then use these rubrics to discuss the survey documentation of several popular surveys originating in the U.S., U.K. and Europe.

This paper is accompanied by a website, where applied researchers can paste example code from SAS, Stata and R and generate corresponding code in other software packages to facilitate the correct use of weights in future. Please visit https://statstas.shinyapps.io/svysettings/ for details.

2. Survey settings in statistical software

The most common public use data file specification of an area probability sampling design is that of a two-stage stratified clustered sample. It is nearly always an approximation to the true sampling design, as most typically the design would include more stages, and some additional modifications of the sampling design variables would be undertaken: true sampling strata or units would be combined or split, units would be swapped with one another, etc., typically in order to mask the true geographical locations of respondents, as geography is one of the strongest factors putting individuals at risk of identification and disclosure (Heeringa et al., 2017, Chapter 4).

In the examples of software syntax below, we rely on the following vignettes:

- A "public use" stratified two-stage design:
  - the data file in the package native format is `PUMS_svy`, with an appropriate extension

  - strata are `thisStrat`
  - clusters are `thisPSU`
  - weights are `thisWeight`
  - Taylor Series Linearization (TSL) is the default variance estimation procedure in these settings
- A "dual frame RDD" design, approximated by an unequal probability design, with replicate weights:
  - the data file in the package native format is `RDD_svy`, with an appropriate extension
  - weights are `thisWeight`
- A design with bootstrap replicate weights:
  - the data file in the package native format is `BSTRAP_svy`, with an appropriate extension
  - the main weights are `thisWeight`
  - the replicate weights are `bsWeight1, bsWeight2, ..., bsWeight100`

In addition, three analyses are discussed: estimation of the total of a continuous variable `y`; cross-tabulation of two categorical variables `sex` and `race`; analysis in subpopulation/domain defined by age restriction: `age` between 18 and 30.

2.1 R

R (R Core Team 2019) is a free, open-source software environment for statistical computing and graphics. The base R provides the computational background and a minimal set of statistical computing (e.g., distributions), while most of the functionality exists in third party packages. Implementation of complex sample survey estimation in `library(survey)` (Lumley 2010) separates the steps of declaring the sampling design and running estimation.

(In terms of reading the input data, we assume that the user follows the best practices of workflow management and uses `library(here)` to identify the root of the project; see Bryan (2017)).

2.1.1 Public use stratified two-stage design

```
# prerequisites
library(survey)
library(here)
# (0) read the data and specify the design
thisSurvey <- readData(here("data", "PUMS_svy.Rdata"))
thisDesign <- svydesign(id =~ thisPSU, strat =~ thisStrat,
      weights =~thisWeight, data =~ thisSurvey)
# (1) estimate the total
(total_y <- svytotal(~y, design = thisDesign) )
# (2) tabulate
(tab1_sex_race <- svymean( ~interaction(sex,race,drop=TRUE),
      design = thisDesign ) )
(tab2_sex_race <- svytable( ~sex+race, design = thisDesign) )
(tab3_sex_race <- svyby(~sex, by = ~race, design = thisDesign, FUN = svymean)
# (3) subpopulation estimation: redeclare the design
young_adults <- subset( design = thisDesign, ( (age>=18) & (age<=30) ) )
(total_y_young <- svytotal(~y, design = young_adults )
```

In the above, the line `( object <- function_call(input1, ... ) )` simultaneously creates and assigns the object, and prints it. Lumley (2010) notes that by default, all functions give missing values (`NA`) when they encounter item missing data. To discard the missing data from analysis, `na.rm=TRUE` should be specified as an option to the `svy...(...,na.rm=TRUE)` functions, with the effect of treating the non-missing data data as a subpopulation. More complex analysis, such as linear models are available through the `survey` package.

2.1.2 RDD unequal weights design

```
# prerequisites
library(survey)
library(here)
# (0) read the data and specify the design
thisSurvey <- readData(here("data","RDD_svy.Rdata"))
thisDesign <- svrepdesign(id =~ 1, weights =~thisWeight, data =~ thisSurvey)
# estimation can use the same syntax as above
svytotal(~y, design = thisDesign)
svymean( ~interaction(sex,race,drop=TRUE), design = thisDesign ) )
young_adults <- subset( design = thisDesign, ( (age>=18) & (age<=30) ) )
svytotal(~y, design = young_adults )
```

Same calls as in the previous block of code, namely analyses specified under (1), (2) and (3), can be used. The `survey` package abstracts estimation details and provides a unified interface that is generally design-agnostic.

2.1.3 The replicate weight design

```
# prerequisites
library(survey)
library(here)
# read the data
thisSurvey <- readData(here("data", "BSTRAP_svy.Rdata"))
# specify the design
thisDesign <- svydesign(weights =~thisWeight, data =~ thisSurvey,
                        repweights =~ "bsWeight[0-9]+", type="bootstrap",
                        combined.weights = TRUE)
# estimation can use the same syntax as above
svytotal(~y, design = thisDesign)
svymean( ~interaction(sex,race,drop=TRUE), design = thisDesign ) )
young_adults <- subset( design = thisDesign, ( (age>=18) & (age<=30) ) )
svytotal(~y, design = young_adults )
```

In the above syntax, `"bsWeight[0-9]+"` is a *regular expression*[1] which, in this case, builds a filter for variable names as follows: 1. must start with the text `bsWeight` exactly; 2. this prefix must be followed by a digit, specified as `[0-9]` 3. this digit must happen at least once, and may happen an unlimited number of times (`+` modifier).

For more examples, see Thomas Lumley's documentation of the `library(survey)` package, as well as Lumley (2010).

[1] See https://regexr.com/

An alternative package is `library(ReGenesees)`. It is not as easily accessible and as regularly updated as the survey package.

2.2 Stata

Stata (StataCorp 2019) is a commercial package that provides most of the functionality through the official release, but also provides ways for the third party developers to code their commands that are indistinguishable from the native Stata commands, at least by syntax. In Stata, survey settings can be specified once with `svyset` command, and be used later with the `svy:` estimation prefix. The settings can be saved with the data set, so that the end users do not have to make this step on their end. This is a recommended best practice for data providers.

2.2.1 Public use stratified two-stage design

```
use data/PUMS_svy, clear
svyset
* if empty, specify svyset on your own
svyset thisPSU [pw=thisWeight], strata(thisStrat)
* estimate the total
svy :  total y
* tabulate
svy : tab sex race, col se
* subpopulation estimation: subpop option
svy , subpop( if inrange(age,18,30) ) : total y
```

2.2.2 The RDD unequal weights design:

```
use data/RDD_svy, clear
svyset
* if empty, specify svyset on your own
svyset thisPSU [pw=thisWeight]
* estimation commands as before
* estimate the total
svy :  total y
* tabulate
svy : tab sex race, col se
* subpopulation estimation: subpop option
svy , subpop( if inrange(age,18,30) ) : total y
```

The estimation commands themselves are identical to those for the cluster+strata designs. Just like in R, the estimation interface is independent of the survey specification interface.

2.2.3 The replicate weight design

```
use data/BSTRAP_svy, clear
svyset
* if empty, specify svyset on your own
svyset [pw=thisWeight], vce(bootstrap) bsrw( bsWeight* ) mse
* estimate the total
svy :  total y
* tabulate
svy : tab sex race, col se
* subpopulation estimation: subpop option
svy , subpop( if inrange(age,18,30) ) : total y
```

The estimation commands themselves are identical to those for the cluster+strata designs.

The `mse` option of the `svyset` command requests the MSE version of the estimator where the original estimate is subtracted, vs. the variance version where the mean of the pseudo-values is substracted when the squared differences are formed.

Starting with Stata 15.1, calibrated weights are supported (Valliant & Dever 2018).

2.3 SAS®

SAS software (SAS Institute 2019) is a commercial statistical package. Nearly all of statistical functionality is implemented via procedures (`PROC`) developed by SAS Institute. In SAS software, survey settings need to be declared in every `SURVEY` procedure. Paths to the data files are declared externally with `libname` statements (not shown).

2.3.1 Public use stratified two-stage design

Estimation of the total (`sum` option of `SURVEYMEANS`):

```
PROC SURVEYMEANS data=thisSurveyLib.PUMS_svy sum;
   WEIGHTS thisWeight;
   CLUSTER thisPSU;
   STRATA thisStrat;
   VAR y;
RUN;
```

Tabulations and cross-tabulations:

```
PROC SURVEYFREQ data=thisSurveyLib.PUMS_svy;
   WEIGHTS thisWeight;
   CLUSTER thisPSU;
   STRATA thisStrat;
   TABLES sex*race;
RUN;
```

Subpopulation analysis:

```
DATA thisSurveyLib.PUMS_svy;
   SET thisSurveyLib.PUMS_svy;
   age_18to30 = (age>=18) & (age<=30);
RUN;
PROC SURVEYFREQ data=thisSurveyLib.PUMS_svy;
   WEIGHTS thisWeight;
   CLUSTER thisPSU;
   STRATA thisStrat;
   TABLES sex*race;
   DOMAIN age_18to30;
RUN;
```

2.3.2 Unequal weights design

Estimation of the total:

```
PROC SURVEYMEANS data=thisSurveyLib.RDD_svy sum;
   WEIGHTS thisWeight;
   VAR y;
RUN;
```

Tabulations and cross-tabulations:

```
PROC SURVEYFREQ data=thisSurveyLib.RDD_svy;
   WEIGHTS thisWeight;
   TABLES sex*race;
RUN;
```

Subpopulation analysis:

```
DATA thisSurveyLib.RDD_svy;
   SET thisSurveyLib.RDD_svy;
   age_18to30 = (age>=18) & (age<=30);
RUN;
PROC SURVEYFREQ data=thisSurveyLib.RDD_svy;
   WEIGHTS thisWeight;
   TABLES sex*race;
   DOMAIN age_18to30;
RUN;
```

2.3.3 Replicate weight design

Estimation of the total (`sum` option of `SURVEYMEANS`): since the algebra of the bootstrap weight computation is identical to that of balanced repeated replication (BRR), the BRR design can be specified as a shortcut (Phillips 2004) with the `varmethod=BRR` option:

```
PROC SURVEYMEANS data=thisSurveyLib.BSTRAP_svy sum varmethod=BRR;
   WEIGHTS thisWeight;
   REPWEIGHTS bsWeight1 – bsWeight100;
   VAR y;
RUN;
```

Tabulations and cross-tabulations:

```
PROC SURVEYFREQ data=thisSurveyLib.BSTRAP_svy varmethod=BRR;
   WEIGHTS thisWeight;
   REPWEIGHTS bsWeight1 – bsWeight100;
   TABLES sex*race;
RUN;
```

Subpopulation analysis:

```
DATA thisSurveyLib.BSTRAP_svy;
   SET thisSurveyLib.BSTRAP_svy;
   age_18to30 = (age>=18) & (age<=30);
RUN;
PROC SURVEYFREQ data=thisSurveyLib.BSTRAP_svy varmethod=BRR;
   WEIGHTS thisWeight;
   REPWEIGHTS bsWeight1 – bsWeight100;
   TABLES sex*race;
   DOMAIN age_18to30;
RUN;
```

A combination of `WEIGHT` and `REPWEIGHTS` produces an MSE version of the estimator. To obtain a variance version (i.e., subtracting the mean of the replicate pseudovalues rather than the estimate based on the main weight), omit the `WEIGHT` statement.

3. Documentation on appropriate design-based analysis techniques for complex sample survey data: rubrics

Large scale data collections are nowadays routinely released to the public. They typically include anonymized, public use survey microdata, along with some variables that include details about the fieldwork itself, and one or several weighting variables that allow any data user to correct for unequal sampling probabilities introduced in the survey design, as well as noncoverage and nonresponse errors. The survey datasets are accompanied with survey documentation that purports to explain the design of the survey and detail the measurements taken. In this section we propose a short checklist to assess quality of survey documentation concerning survey design features specification in software. We intend this checklist to be used primarily by those producing survey data and its documentation, so that these organizations could make sure their data products are sufficiently user-friendly.

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** This would be a person with training on par with or exceeding the level of the Lohr (2010) or Kish (1965) textbooks, and applied experience on par with or exceeding the Lumley (2010) or Heeringa, West and Berglund (2017) books.
2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** This would be a person who has some background / training in applied statistical analysis, but has only cursory knowledge of survey methodology, based on at most several hours of classroom instruction in their discipline "methods" or “metrics” class, or a short course at a conference.
3. **Is everything described succinctly in one place, or scattered throughout the document?** It is of course easier on the user when all the relevant information is easily available in a single section. However, some reports put information about weights in one place, e.g. where sampling was described, while information about other complex sampling features (e.g., cluster/strata/variance estimation) only appears some twenty pages further.
4. **Are examples of specific syntax to specify survey settings provided?** Has the data producer provided worked and clearly-annotated examples of analyses of the complex sample survey data produced by a given survey using the syntax for existing procedures in one or more common statistical software packages? And as a bonus, have examples been provided in multiple languages (e.g., SAS, R, and Stata)?
5. **Are there examples given for how to answer substantive research questions?** In all languages, there are specific ways to run commands that are survey-design-aware. In other words, only specifying the design may not be sufficient in ensuring that estimation is done correctly. For instance, are examples provided for both descriptive and analytic (i.e., regression-driven) research questions?
6. (Bonus) **Is an executive summary description of the sampling design available?** Many researchers would appreciate a two-three sentence paragraph to summarize the sampling design that they could copy and paste into their papers, e.g.,

   > *{This survey} is a three-stage area sampling design survey with census tracts, households, and individuals as sampling units. The*

> *final analysis weights provided by {the organization who collected the data} account for unequal selection probabilities, nonresponse, and study eligibility, and are used in all analyses reported in this paper. Standard errors are estimated using the complex survey bootstrap variance estimation procedures.*

or

> *{This survey} is a dual-frame RDD survey that collected data on both landline and mobile phones. The final analysis weights provided by {the organization who collected the data} account for unequal selection probabilities, nonresponse, and study eligibility, and are used in all analyses reported in this paper. Standard errors are estimated using Taylor series linearization, the default analytical method available in most statistical packages.*

7. (Bonus) **What kinds of references are provided?** It is often helpful to the end users if the description of the sampling design features is accompanied by the references to (a) methodological literature describing them in general, and (b) technical publications specific to the study in question, such as the JSM or AAPOR proceedings, technical reports on the provider website, or publications in technical literature describing the study, if appropriate. For example, the description of clustered sampling designs used in the U.S. Census Bureau large scale surveys such as the American Community Survey or Current Population Survey could refer to general descriptions of stratified clustered surveys, to the user Handbooks (Census Bureau 2009), and to the technical papers on variance estimation (Ash 2011).

We now use the seven rubrics defined above to "score" several existing examples of documentation for public-use survey data files based on these criteria. For example, if the documentation for a public-use data file successfully satisfies / meets the first five rubrics above, the documentation will be scored 5/5. These scores are designed to be **illustrative**, in terms of rating existing examples of documentation for public-use data files on how effectively they convey complex sampling features and how they should be employed in analysis to users. The scores are designed to motivate data producers to improve the clarity of their documentation for a variety of data users hoping to analyze large (and usually publically-funded) survey data sets.

3.1 Practical strategies to dealing with existing documentation

When asked to analyze an existing data set that features complex survey data, we typically rely on a number of heuristics to figure out what the survey statisticians intend for the ultimate users to do.

1. Search the documentation for the software footprints as keywords: `svyset` per Stata, `PROC SURVEY` per SAS, `svydesign` per R `library(survey)`.
2. If that fails, search for "sampling weight", "final weight", "analysis weight", "survey weight" or "design weight". You can search for "weight" per se but you should expect that this is likely to produce many false positives (weight as a physical measurement in kg), especially in health studies.
3. See if there is any description of the sampling strata and clusters near the text where weights are mentioned.

4. Search for "*PSU*" and "*cluster*" and "*strata*" and "*stratification*" to find the variables that needed to be specified in survey settings.
5. Search for "*variance estimation*", the generic technical term to deal with complexities of survey estimation.
6. Search for "*replicate weights*", "*BRR*", "*jackknife*" and "*bootstrap*", the keywords for the popular replicate variance estimation methods.

In reviewing weighting documentation of existing surveys, we have also encountered more obscure language such as “pseudovalues”, “pseudostrata”, “pseudounits”, “variance replicates”, “variance units”, “pseudoreplicates” and some other terms indicating that the variables provided for variance estimation may not be the true sampling design variables. While technically correct, such language does little to help an inexperienced user in identifying the relevant settings to be applied, primarily through disconnect between the “textbook” terms and the terms used in documentation.

4. Evaluating documentation in practice

In this section, we will evaluate a convenience sample of the documentation for several public use survey data files (PUFs). The goal of this section is not to provide our overall assessment of weighting procedures across all datasets; we merely want to illustrate how several large-scale and much used survey datasets have described what was done in their complex sampling designs and corrections. We will apply the above rubrics to see how the documentation compares in terms of effectively describing appropriate analysis techniques to data users. Additional examples, including those with lower ratings, are available at the main project webpage, https://github.com/skolenik/svyset_manifesto.

4.1 The National Survey of Family Growth (NSFG), 2013--2015

Rating: ⭑⭑⭑⭑⭑

Funding:

- Eunice Kennedy Shriver National Institute of Child Health and Human Development
- Office of Population Affairs
- NCHS, CDC
- Division of HIV/AIDS Prevention, CDC
- Division of Sexually Transmitted Disease Prevention, CDC
- Division of Reproductive Health, CDC
- Division of Birth Defects and Developmental Disabilities, CDC
- Division of Cancer Prevention and Control, CDC
- Children’s Bureau, Administration for Children and Families (ACF)
- Office of Planning, Research and Evaluation, ACF

Data collection: The University of Michigan Survey Research Center (http://src.isr.umich.edu)

Host: The National Center for Health Statistics (http://www.cdc.gov/nchs/)

URL: http://www.cdc.gov/nchs/nsfg

Rubrics:

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** Yes. Electronic documents like Example 1: Variance Estimates for Percentages linked from the documentation page under *Variance estimation* subtitle make it very easy for survey statisticians and applied researchers alike to correctly declare complex sampling features to survey analysis software for design-based analyses.
2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** Yes. See above.
3. **Is everything that the data user needs to know about the complex sampling contained in one place?** Yes, although very little (if anything) is said about the actual complex sampling design. Instead this information appears in separate electronic files, such as Sample Design Documentation. This is out of necessity, however, given the complexity of the NSFG sampling design, and all of the information that a user needs to compute weighted point estimates and estimate variance accounting for the complex sampling can be found in examples like the one indicated above.
4. **Are examples of specific syntax for performing correct design-based analyses provided?** Yes. Three examples are clearly documented (tabulations for categorical variables; means for continuous variables; analysis with domains/subpopulations) and linked on the main documentation page, and both syntax and output are included in each case. Bonus: syntax and output are provided for both SAS and Stata.
5. **Are examples of analyses need for addressing specific substantive questions provided?** Yes; see previous item.
6. **(Bonus) Is an executive summary of the sampling design provided?** Yes; such an executive summary is given in the first section of the main sample document
7. **(Bonus) What kinds of references are provided?** There are several references to the most important sampling design literature included in Section 11 of the document linked above.

Score: 5++/5

The NSFG provides an excellent example of the type of documentation that needs to be provided to data users to minimize the risk of analytic error due to a failure to account for complex sampling features.

Accessed on 2018-07-15. Disclaimer: one of the authors of this paper, Brady West, is an investigator on NSFG. Other authors of the paper believe that the high quality of NSFG documentation owes, at least to some extent, to Dr. West's involvement.

4.2 The Population Assessment of Tobacco and Health

Rating: ☆☆☆☆☆

Funding: The Population Assessment of Tobacco and Health (PATH) Study is a collaboration between the National Institute on Drug Abuse (NIDA), National Institutes of Health (NIH), and the Center for Tobacco Products (CTP), Food and Drug Administration (FDA).

Data collection: Westat (http://www.westat.com)

Host: The National Addiction and HIV Data Archive Program

URL: https://www.icpsr.umich.edu/icpsrweb/NAHDAP/series/606

Rubrics:

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** Yes. Section 5 of the Public-Use Files User Guide provides clear detail

on the calculation and names of the various weight variables that can be used for estimation. This section also discusses variance estimation, and clearly describes the replicate weights that have been prepared for data users enabling variance estimation. Software options are also discussed in this section, and code illustrating the use of multiple programs for the protype example analyses is provided in Appendix A.

2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** Yes. Appendix A of the User Guide is very helpful, given that it provides annotated example code for several different packages. Section 5 is aimed at survey statisticians, and will be overwhelming to an audience that is less technically prepared.
3. **Is everything that the data user needs to know about the complex sampling contained in one place?** Yes; Section 5 provides all of the necessary sampling information for analysis purposes, and Appendix A contains all of the necessary code for actual practice.
4. **Are examples of specific syntax for performing correct design-based analyses provided?** Yes. Appendix A of the Public-Use Files User Guide is an excellent example of providing this kind of resource for data users.
5. **Are examples of analyses needed for addressing specific substantive questions provided?** Yes. Appendix A illustrates a variety of potential analyses that data users could perform.
6. **(Bonus) Is an executive summary of the sampling design provided?** Chapter 2 of the User Guide provides a detailed summary of the sampling design, which serves as an executive summary.
7. **(Bonus) What kinds of references are provided?** There are several references to the most important sampling design literature included at the end of the User Guide.

Score: 5++/5

The PATH PUF user guide is another excellent, gold-standard example of detailed and useful information designed to make the life of the survey data user easier.

Accessed on 2018-12-17.

4.3 European Social Survey Rounds 1-7

ESS represents an interesting example of a survey on which we had observed tangible improvements in documentation through the lifetime of our project. We provide both the early scoring, and the later one based on redesigned

Funding: European Commission, Horizon 2020. Rounds 1-7 of ESS have been founded by national science foundations and/or European national governments.

Data collection: coordinated by City University, London, UK. Data collection in separate European Countries coordinated within every country.

Host: European Social Survey, formerly at Norwegian data Archive

URL: www.europeansocialsurvey.org

4.3.1 Early assessment as of July 2017

Rating: ☆ ☆

Weighting documentation:
http://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf

Rubrics:

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** Yes. The European Social Survey is a repeated cross-sectional study conducted in about 30 different countries in Europe. Sampling is conducted within every country, using either listing methods or registers (of individuals or addresses). Three weights (design, poststratification and population equivalence weights) are included in the main datafile. This allows for Horvitz-Thompson estimation, but not the specification of a complex survey design. However, an Integrated Sample data file does include information on stratification or cluster variables, as well as selection probabilities for every respondent. On top of this, a multilevel file adds regional indicators to the main datafile, allowing for multilevel-analysis
2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** Yes, three weights are provided: a design weight, a poststratification weight and a population equivalence weight. Guidance is included on how to combine the three weights, and when to use what weight in some examples of analyses.
3. **Is everything that the data user needs to know about the complex sampling contained in one place?** Documentation is scattered across many different documents and files on the ESS website. However, most users in practice would use one round of ESS. In that case, the country report files contain details on how fieldwork (including sampling) was conducted. One good aspect of the European Social Survey is that the users are explicitly warned that data need to be weighted when data are downloaded from the ESS website. However, there isn't an accompanying warning about using the sampling design variables for variance estimation as well.
4. **Are examples of specific syntax for performing correct design-based analyses provided?** No.
5. **Are examples of analyses need for addressing specific substantive questions provided?** There are a few examples of data management code, but not of the complex survey analysis syntax.
6. **(Bonus) Is an executive summary of the sampling design provided?** There is an executive summary that describes the basic sampling methodology. There is no easily accessible executive summary that explains how and why sampling differs over the countries.
7. **(Bonus) What kinds of references are provided?** There are references to standard textbooks on complex survey design, and references to other documents on the ESS website, with more detailed documentation.

Score: 2/5

The ESS is a typical example of documentation written by survey statisticians for survey statisticians, and it takes a survey statistician to process it and come up with the requisite syntax. Novice users may be deterred by the complexity of documentation, and would choose to either underutilize the resource, or would otherwise have to ask the survey providers additional questions, especially in the case where country comparative analyses are conducted.

Accessed on 2017-07-19.

4.3.2 Continued assessment as of September 2019

Rating: ⭑ ⭑ ⭑ ⭑

Weighting documentation:
http://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf

Round 8 User Guide:
https://www.europeansocialsurvey.org/docs/round8/methods/ESS8_sddf_user_guide_1_1.pdf

Rubrics:

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** Yes. The European Social Survey is a repeated cross-sectional study conducted in about 30 different countries in Europe. Sampling is conducted within every country, using either listing methods or registers (of individuals or addresses). Three weights (design, poststratification and population equivalence weights) are included in the main data file. This allows for Horvitz-Thompson estimation, but not the specification of a complex survey design. However, an Integrated Sample data file does include information on stratification or cluster variables, as well as selection probabilities for every respondent. On top of this, a multilevel file adds regional indicators to the main datafile, allowing for multilevel-analysis
2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** Yes, three weights are provided: a design weight, a poststratification weight and a population equivalence weight. Guidance is included on how to combine the three weights, and when to use what weight in some examples of analyses. ESS Round 8 documentation discusses the sampling design variables such as strata and clusters.
3. **Is everything that the data user needs to know about the complex sampling contained in one place?** Documentation is scattered across many different documents and files on the ESS website. One good aspect of the European Social Survey is that the users are explicitly warned that data need to be weighted when data are downloaded from the ESS website. Round 8 User Guide does compile the description of all the design variables. It is unclear whether users of other rounds will stumble upon it.
4. **Are examples of specific syntax for performing correct design-based analyses provided?** Yes. Box 2 in Section 3.2 "Estimating standard errors" of the Round 8 User Guide provides Stata `svyset` syntax.
5. **Are examples of analyses need for addressing specific substantive questions provided?** Yes. Box 3 in Section 3.2 "Estimating standard errors" of the Round 8 User Guide provides Stata syntax to obtain design-adjusted estimates, however the syntax is incorrect as it uses subsetting the data rather than subpopulation/domain estimation (West, Berglund and Heeringa 2008). The subsequent discussion of the differences between naïve and design-adjusted estimates is very helpful.
6. **(Bonus) Is an executive summary of the sampling design provided?** There is an executive summary that describes the basic sampling methodology. There is no easily accessible executive summary that explains how and why sampling differs over the countries.

7. **(Bonus) What kinds of references are provided?** There are references to standard textbooks on complex survey design, and references to other documents on the ESS website, with more detailed documentation.

Score: 4/5

The ESS provides a mix of legacy documentation written by survey statisticians for survey statisticians, and the more recent documentation aimed at the non-statistical users. The use of multi-country, multi-round data sets remains very complex.

Accessed on 2019-09-16.

4.4 A Portrait of Jewish Americans

Rating: ☆ ☆ ☆ ☆

Funding: The Pew Research Center's 2013 survey of U.S. Jews was conducted by the center's Religion & Public Life Project with generous funding from The Pew Charitable Trusts and the Neubauer Family Foundation.

Data collection: Abt SRBI under contract to Pew Research Center

Host: Pew Research Center http://www.pewresearch.org/

URL: http://www.pewforum.org/dataset/a-portrait-of-jewish-americans/

Rubrics:

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** Yes; survey documentation explains the differences between the household and the person-level weights, and stresses that the bootstrap weights should be used for variance estimation.
2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** Yes; Stata syntax is provided early in the document, or can be found by search in the PDF file.
3. **Is everything described succinctly in one place, or scattered throughout the document?** Yes; all of the relevant information is contained in the **Key Elements of the Data** section in about 2 pages.
4. **Are examples of specific syntax to specify survey settings provided?** Yes; item 6 of **Key Elements of the Data** section identifies the variables and provides Stata syntax for individual level and household level analyses. (Search for any of `Stata`, `SAS`, `weight`, `svyset` would lead the researcher to this information.) A warning is given that SPSS Statistics Base package cannot correctly compute standard errors.
5. **Are there examples given for how to answer substantive research questions?** No examples are given.
6. (Bonus) **Is an executive summary description of the sampling design available?** Sampling design is described in painstaking detail in about 9 pages. No short summary of the design is available from the technical documentation, although such a summary can be found in the substantive report (Pew Research Center 2013).
7. (Bonus) **What kinds of references are provided?** No additional references are given.

Score: 4+/5

A Portrait of Jewish Americans is a very well described survey that most researchers will be able to analyze correctly by following the instructions of the data provider. Slight limitations of the documentation is that examples of the settings are only given for one package, Stata, and no examples of substantive analyses, e.g. those leading to the headline tables in the substantive report, are provided.

Accessed on 2018-12-11.

References

Ash, S. (2011). Using Successive Difference Replication for Estimating Variances. *Proceedings of the Survey Research Methods Section*, 3534–3548, American Statistical Association, Alexandria, VA.

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review*, 78 (2), 161–188. https://doi.org/10.1111/j.1751-5823.2010.00112.x

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality* (Vol. 335). John Wiley & Sons.

Bryan, J. (2017) Project-oriented workflow. Technical report available at https://www.tidyverse.org/articles/2017/12/workflow-vs-script/

Census Bureau (2009). *What Researchers Need to Know: ACS Handbook*. Technical Report. https://www.census.gov/library/publications/2009/acs/researchers.html

Groves, R. M. (2004). *Survey errors and survey costs* (Vol. 536). John Wiley & Sons.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (Vol. 561). John Wiley & Sons.

Groves, R., and Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. Public Opinion Quarterly, 74 (5), 849–879. https://doi.org/10.1093/poq/nfq065

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis*. Chapman and Hall/CRC.

Kalton, G., and Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics*, 19 (2), 81–97.

Khera, R., et al (2017). Adherence to Methodological Standards in Research Using the National Inpatient Sample. Journal of the American Medical Association, DOI: 10.1001/jama.2017.17653

Kish, L. (1965) *Survey sampling*. New York: Wiley.

Kolenikov, S. (2016). Post-stratification or non-response adjustment? *Survey Practice*, 9 (3). https://doi.org/10.29115/SP-2016-0014

Lohr, S. L. (2010). *Sampling: Design and Analysis: Design and Analysis*. Chapman and Hall/CRC.

Lumley T (2010).*Complex Surveys: A Guide to Analysis using R*. John Wiley & Sons, Hoboken, New Jersey.

Pew Research Center (2013). A Portrait of Jewish Americans. Technical report, available at http://www.pewforum.org/2013/10/01/jewish-american-beliefs-attitudes-culture-survey/

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

SAS Institute (2019). Base SAS 9.4; SAS/STAT 14.1 User Guide. Cary, NC.

StataCorp, L. P. (2019). Stata statistical software: Release 16. *College Station TX*.

Valliant, R., & Dever, J. A. (2018). *Survey weights: a step-by-step guide to calculation*. College Station, TX: Stata Press.

Vallian, R., Dever, J., & Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York.

West, B. T., Berglund, P., & Heeringa, S. G. (2008). A Closer Examination of Subpopulation Analysis of Complex-Sample Survey Data. *The Stata Journal*, 8 (4), 520–531. https://doi.org/10.1177%2F1536867X0800800404

West, B. T., Sakshaug, J. W., & Aurelien, G. A. S. (2016). How big of a problem is analytic error in secondary analyses of survey data?. *PloS one*, *11*(6), e0158120.