Statistical inference with weights and survey design variables

A practical guide using UKDS datasets

Pierre Walthéry and Jennifer Buckley

08 August 2023

Introduction

This note aims at discussing and setting out guidelines for population inference using weights and survey design variables. It focuses on a limited number of practical procedures and does not discuss the theoretical underpinnings of survey design, sampling or estimation with weighted survey data. The content is based on technical documents by data producers such as the Office for National Statistics as well as the relevant statistical literature. Examples are currently drawn from the Labour Force Survey, the Family Expenditure Survey and the British Social Attitudes Survey and will be gradually be expanded. A list of key references and online tuturials is provided in the bibliography.

1. Survey design and weights variables

Social surveys enable researchers and analysts to learn about the characteristics of human populations. This is achieved by way of conducting statistical inference, the process through which unknown characteristics (sometimes called parameters) of 'large' populations are estimated with the help of smaller samples that are drawn from them. Estimation of population parameters traditionally consist in computing two pieces of information: a measure of the likely typical value of interest also known as point estimate, together with an indication of its uncertainty or precision.

When certain conditions are met, such as when sample members are randomly selected and sample size is large enough, surveys and the parameters inferred from them are representative of the corresponding population [Loi des grands npombres]. However, in part as a result of design decisions, in part – and increasingly so – due to non response, uncorrected population estimates from real-life social surveys present some degree of bias.

Robust, that is unbiased estimates need to be at the same time 'representative' – reflect the characteristic of interest in the population, but also precise enough for the inference to be meaningful. The latter issue is the most complex to deal with in the context of statistical inference. It is usually considered that in order to produce robust population estimates, including as much of the survey design information as possible alongside (non response and sampling) weights is

required. Conversely, estimates computed without weights or accounting for survey design will at best present some degree of bias be altogether unreliable.

Computing weighted estimates and accounting for survey design usually require specific procedures that are not usually very well documented as the relevant statistical techniques are more complex and overlooked in introductory textbooks, and their practical implementation in statistical software may not always be clear. It is therefore necessary to add some clarity to this situation and provide adequate guidelines in order for UKDS users to properly implement robust estimation strategies that are adapted to their needs.

Note possibly add existing introductory content on survey design here

Survey weights

Weights are a special type of numeric variables included in survey datasets, whose value tends to be the inverse of the relative 'importance' of sampled observations. They are designed to prevent estimates from being biased, that is reflecting a value that is not representative of the population. They are usually made of at least two components:

- a *design* component that accounts for issues of unequal probability of selection of sample members resulting from survey design;
- a *non-response* component, correcting for (known) lower propensity to take part to surveys among certain categories of respondents.

Rather confusingly, these components are sometimes labelled 'weights' in their own right, even if in practice they are most of the time merged into a single variable.

Survey weights may also be rescaled in order to inflate sample counts to population totals thus becoming *grossing weights* which enable estimating populations size. In that sense the numerical values of the weights attached to observations are an indication of the number of units these observation 'represent' in the population.

Computation of weights rely on calibration algorithms that optimise the conditional distribution of the weighting variables given the sample size (for example the conditional distribution of people by age, gender and economic status) with a view to strike a balance between minimising standard errors and maximising representativeness.

Survey design

Contrarily to common sense assumptions, collecting data about people at random is far from being straightforward. There is no such thing as a list of all UK residents to pick from, and even if there were one, some people who are selected are less likely to take part to survey than others.

As a result most UK social surveys rely on sampling techniques such as multi-stage clustering and stratification, alongside sampling proportional to size in order to strike a compromise between tackling non response and unequal probability of selection, improve representation of hard to reach groups whilst keeping fieldwork costs down.

Clustering involves dividing the population into groups (better thought of as 'mini populations') and randomly select those from which final sampled units will be drawn as a subsequent stage. Stratification consist of grouping the population according to characteristics known to be associated with non response, and randomly draw sample members from each strata, sometimes disproportionately. Typically, estimates from clustered surveys yield less precise estimate than non clustered ones, and conversely stratified surveys, more precise than non stratified ones.

Survey design variables typically consist of identifiers for the strata and/or clusters used when, especially the Primary Sampling Units (PSU) used during the sampling process. Used in conjunction with weights, they enable researchers to produce more accurate estimates.

2. Things to keep in mind when analysing survey data

The variability (and therefore the degree of precision with which they can be estimated) of point estimates is contingent on survey weights and survey design. Although therefore the optimum approach to estimating population parameters from surveys relies on using both weights and survey design variables, it is not always possible to go down that path. In practice, trade-offs have to be made depending on several factors. Let us briefly consider them.

Data availability

Most UKDS datasets are available under *End User License*. This presents the advantage of enabling large numbers of users to access data with a minimal level of formalities to go through but often comes at the significant cost that survey design variables are not included by the data producer, due to concerns about the risk of personal information disclosure. There are notable exceptions, such as for example the British Social Attitudes survey which does include survey design variables in some of its releases.

For a number of key studies such as the Labour Force Survey or the Family Resources Survey, users may apply for access to a version of the data that includes survey design information via the (virtual) SecureLab or at the UKDS Safe Room. Application for access to these facilities can be a lengthy process, and not practically feasible for all researchers, in particular those outside academia or large organisations. More information is available on the UKDS website. There are also a large number of studies for which such controlled access is not available. The consequence is that in a significant number of cases, there will inevitably be a ceiling in the level of precision of the estimates most will be able toreach.

Sensitivity of the analysis

Not all analyses necessarily require the highest degree of precision. Reflecting on the stake of their intended analysis will help users decide how important it is to strive to use the most robust estimation technique available or instead to settle for one that is 'good enough'. Typical usages of survey data could be seen as lying on a continuum ranging from 'playing with the data' to producing numbers that will be subject to public scrutiny, or that will be used in policymaking. The latter require such a degree of precision – for example when publishing official population estimates or writing a research article, other less so – for instance when exploring data or preparing examples for

teaching. In the former cases, users may simply need to get a rough idea of a population estimate or the interval within which it may lie.

Complexity of the analysis

What an analysis actually entails will help determine whether accessing survey design variables is crucial or not. Estimation involving small numbers of observations will be more at risk of providing incorrect estimates if survey design variables are not taken into account. Similarly, interest for specific subgroups of the population (also known as domains) rather than the population as a whole will involve more complex estimation techniques as domain estimation needs to account for the distribution os weights as whole, not just those of the subgroup of interest.

These analytical scenarios could be seen as lying on a continuum ranging from producing simple univariate descriptive estimates for the population as a whole to complex estimation of small groups characteristics and/or multivariate analysis. The former is conceptually and practically more straightforward than the latter. In some cases the estimates of interest may already have been published by the data producer using the adequate estimation technique. Data producers may also have published *design factors* ie numbers allowing to adjust the precision of estimates produced without survey design variables. Examples of such design factors for the Labour Force Survey and the Family Resources Surveys are provided below.

Software issues

Because weighting can be used in other contexts than inference from surveys, most statistical software have options for directly weighting estimation commands "on the fly" outside of procedures accounting for survey design (such as the R *Survey* package, the SPSS *Complex Survey* add-on and Stata's *svy:* set of commands).

This can lead some users to solely rely on weighted commands without explicitly declaring the survey design in their analysis which raises issues:

- Whereas weighted commands will most of the time compute the correct point estimates, they will also silently produce biased estimates of their precision (standard errors or confidence intervals), based on the incorrect assumption that the sample was collect via simple random sampling. Depending on the survey design, this will lead to under- or over- estimation of standard errors and confidence intervals, and could affect the validity of statistical tests, in particular if small groups within the population are involved.
- In addition, there are specific cases where estimation of standard errors and confidence intervals will be not just biased but wholly incorrect: the standard (ie command-based) weighting procedure of SPSS ans SAS relies on population rather than sample totals to compute them, which results in unrealistic values.
- Software such as Stata does not allow users to directly compute confidence interval or use sampling weights outside of survey commands. This may lead users to rely on 'quick and dirty' tricks that will help them quickly produce weighted point estimates, with incorrect standard errors.

What are we in fact estimating?

Users can prioritise producing weighted point estimates over estimating their precision and the factors that influence it - chiefly survey design variables. It can be tempting indeed to consider that the goal of statistical inference mainly consists in producing 'representative' point estimates of a quantity of interest such as the 'mean weight of adult males', the 'median poverty rate', or the value of some regression coefficient in a multivariate study with estimates of their precision a secondary consideration, or a qualifier of the point estimate.

This is potentially risky. Point estimates can be at the same time representative *and* imprecise, and therefore carry little practical meaning. It could also be argued that focusing too narrowly on single value population estimates implicitly entertains the idea that such unique, 'true' value exist. As these in fact constantly vary, different surveys will return inevitably different estimates.

Instead, conceiving from the start these two aspects as a single reality – a range of plausible values we think a parameter of interest can take in the population, with a certain degree of confidence – could help alleviate such a risk and most importantly provide a more accurate reflection of the reality we seek to describe. Striving to produce confidence intervals whenever it makes sense to do so will help the notion that precision and therefore inevitably survey design are key to robust estimation.

3. Statistical inference with survey data: practical steps

Ultimately there should be a flowchart here or in the next section

This section provides practical recommendations for robust inference taking into account the factors highlighted above. In general, four strategies are available when conducting population inference from survey data. They are listed in order or recommendation by the UK Data Service.

1. Estimation accounting for weights and survey design using survey-specific commands
2. Estimation accounting for weights only using survey-specific commands
3. Estimation using weighted standard commands
4. (Unweighted estimation)

- *Strategy 1*, using weights alongside sample design variables when conducting statistical inference is the statistically most robust way to compute population estimates with survey data and should be prioritised by users whenever possible. In real life research however, this option is not always available. Accessing survey design variables can prove challenging as they are not always provided by data producers or may require applying for a special version of the data, which may prove time consuming.

- In the absence of survey design information, *Strategy 2* should be considered the second best option. The value of point estimates are likely to be identical to those produced under Strategy 1, but the confidence intervals/standard errors will be biased – ie too narrow or wide depending on the survey design, which should be explicitly mentioned alongside the results. Information from the data documentation should provide information Using survey-specific commands is a recommended option over simply applying weights to standard commands, as it will avoid getting incorrect estimates (SAS and SPSS), is the only option available for

computation with survey weights or obtaining confidence intervals (Stata), or coherent survey data analysis (R).

- It can be understandable that in this context some users privilege *Strategy 3* which tend to focus on producing weighted estimates using standard commands and give little consideration to the methodological implication of this approach. Whereas point estimates are likely to be identical to those produced under Strategy 1 and 2, SAS and SPSS users are likely to produce incorrect confidence intervals/standard errors. R and Stata users might get precision estimates that are close to those produced using Strategy two, but there is not guarantee that this will be the case. Overall UKDS only recommend following this strategy in case of low sensitivity analysis.
- As population estimates produced without weights or survey design variables will almost certainly be unreliable Strategy 4 should be discouraged except when data usage is purely descriptive. For example when teaching non-inferential (ie descriptive) statistical techniques.

3.1 Medium to high sensitivity analysis: workflow

Most of the time survey researchers or data analysts are required to produce a confidence intervals or provide an indication of the degree of precision of their point estimate, usually with standard errors, whose correct estimation as we saw depends on the amount of information held about the survey design.

1. **If survey design variables are available** a typical workflow could involve (see examples in Section 4):

- Finding out about the survey design and identify the relevant weights and survey design variables in the data documentation;
- Declaring the survey design using software-specific commands
- Producing the estimates of interest, using survey design specific estimation commands available
- Documenting the confidence interval for the estimate of interest or alternatively the point estimates *and* its standard error.
- If required, provide a brief discussion of the possible source of bias of the results (specifically under/over estimation of the uncertainty of the estimates)

2. If the survey design variables are not included in the EUL version of the data but are available under controlled access: perform a cost vs benefits analysis of applying for controlled access for instance via the UKDS SecureLab, a process that can take some time. Information about how to apply for Secure Lab Access is available on the UKDS website.

3. If the **survey design variables such as strata, cluster, or primary sampling unit are not available** an alternative workflow could consist in:

- If the user is interested in overall population characteristics, checking whether the estimates of interest may already have been published by the data producer, in which case they may be directly cited instead of computed from data.

- Finding out about the survey design in the data documentation and identify the weights variable ;
- Declaring the survey design as simple random sampling using software-specific commands
- Producing the estimates of interest, using survey design specific estimation commands available
- Checking whether the data producer has published design factors that could be used to remedy to biased confidence intervals/ standards errors computed without survey design variables (for example design factors computed for the same population at another point in time). A design factor is a number by which to multiply standard errors estimated under the assumption of simple random sampling, that will adjust it for survey design characteristics.
- Documenting the resulting confidence interval for the estimate of interest or alternatively the point estimates *and* its standard error.
- If no design factors are available for the estimates of interest, an explicit mention of the likely nature and cause of bias is good practice ie under estimation in case of cluster sampling, over estimation in case of stratified sample, usually available from the survey documentation. The wider the initial confidence interval (ie computed under SRS assumptions) the larger the likely bias. Or from another perspective, the smaller the (sub)sample, the larger the likely bias. In cases of conducting significance testing with small subsample or groups, it would be a good practice to only consider test outcomes significant at P<.01 or p<.001.

4. Computing SDI estimates for subpopulations (also known as 'domains') rather than for the population as a whole requires extra precautions. This is the case for example when we are interested in the mean age by employment status, or some other categories, or alternatively, in analyses restricted to a subset of the population (for example only those in employment). The key differences is that when computing domain estimates we are in fact producing estimates about a group of the population whose size we also need to estimate. This requires ensuring that the whole distribution of weights in the sample is taken into account, not just the weights values for the groups we are interested in. Failure to do so might result in computing incorrect point estimates and standard errors/confidence intervals. SDI commands in statistical software are designed to tackle this potential issue.

3.2 Lower sensitivity analysis

We do not recommend using command-specific or casual weighting for inferential analysis, but there are circumstances where this will be the only option open to users. There are also cases when users are not interested in knowing about the uncertainty of their estimates (ie their confidence interval, standard errors of point estimates, or conduct statistical testing), for example because they are simply learning or teaching basic statistical concepts or how to use software.

In such cases, it can be acceptable to compute point estimates by applying weights to commands that accepts them, without using survey design specific functions. Most of these will provide the correct point estimate. By default however, some statistical software will also provide an estimate of standard errors or confidence intervals, which is likely to be misleading as they 'silently' assume

simple random sampling, and in some cases will carry out computation with population (ie grossed) totals, resulting in the incorrect values.

4. Study-specific weighting and sample design information

4.1 British Social Attitudes Survey

The BSA is a three stage stratified random survey, with postcode sectors, addresses and individuals as the units selected at each stage. Primary sampling units were furthermore stratified according to geographies (sub regions), population density, and proportion of owner-occupiers. Sampling rate was proportional to the size of postcode sectors (ie number of addresses). Some issues of the BSA such as the 2017 include survey design information. The 2017 issue included information about Primary Smapling Units (`Spoint`), strata (`StratID`). Weights are called `WtFactor`.

4.2 Labour Force Survey

The LFS is a geographically stratified random survey. For the main part Primary sampling units are addresses within postcode sectors, drawn from the Small Users Postcode Address File (PAF). The small users PAF is limited to addresses which receive, fewer than 50 items of post per day. In a small number of cases a second stage sampling occurs where several households exist at a given address. A clustering effect is also present to the extent that units of observations are individuals withing households, and that some groups are clustered within these, typically ethnicity. LFS weights: - PWTxx – person level sampling weight; enables inferring population counts - IWTxx - Person-level sampling weight for income analysis (ie subsample of people in paid work) - PHHWTxx - Household-level sampling weight (for household-level analysis)

4.3 Family Resources Survey

The FRS is a stratified clustered random survey, with survey design differing slightly between countries of the UK. In great Britain, Primary sampling units are postcode sectors, drawn from the Small Users Postcode Address File (PAF). The small users PAF is limited to addresses which receive, fewer than 50 items of post per day. Before being selected, PSUs are stratified according to geography, proportion of household reference persons from higher social classes in the area, proportion of economically active respondents in the area, and proportion of economically active men who ware unemployed. In Northern Ireland, the sample is a systematic random sample of addresses.

FRS weights: GROSS4: person-level sampling weight; enables inferring population counts

5. R examples

5.1 SDI Inference with design information using R

The R *Survey* package provides a comprehensive set of function for computing point estimates and reliability from survey data. R does not provide a centralised/unified sets of commands for

computing weighted estimates. Implementation of statistical theory may vary between packages, but algorithms are usually documented in package documentation.

Example 1 Estimating the proportion of people interested in politics using the 2017 British Social Attitudes Survey

```
rm(list=ls())
library(dplyr) ### Data manipulation functions
library(haven) ### Importing stata/SPSS files
library(Hmisc) ### Extra statistical functions
library(survey) ### Survey design functions

bsa17<-read_spss("data/UKDA-8450-spss/spss/spss25/bsa2017_for_ukda.sav")
dim(bsa17)
```

```
## [1] 3988  580
```

Once this is done we can specify the survey design: using `Spoint` as Primary Sampling Unit, `StratID` as strata, and `WtFactor` as weights. R does this by creating a `svydesign` object, ie a SDI version of the data, which will be used for subsequent estimation.

```
bsa17.s<-svydesign(ids=~Spoint, strata=~StratID, weights=~WtFactor,data=bsa17)
class(bsa17.s)
```

```
## [1] "survey.design2" "survey.design"
```

Mean age and its 95% confidence interval

We can now produce a first set of estimates using this information and compare them with those we would have got without accounting for the survey design. We will compute the average (ie mean) age of respondents in the sample. We will need to use `svymean()`

```
svymean(~RAgeE,bsa17.s)
```

```
##         mean     SE
## RAgeE 48.313 0.4236
```

By default `svymean()` computes the standard error of the mean. We need to embed it within `confint()` in order to get a confidence interval.

```
confint(svymean(~RAgeE,bsa17.s)) ### Just the confidence interval...
```

```
##          2.5 %  97.5 %
## RAgeE 47.48289 49.1433
```

```
round(
  c(
    svymean(~RAgeE,bsa17.s),
    confint(svymean(~RAgeE,bsa17.s))
```

```
    ),
  1)### Estimate and CI, rounded
```

```
## RAgeE
##  48.3  47.5  49.1
```

Computing a proportion and its 95% confidence interval

We can now similarly compute the distribution of a categorical variable in the population by estimating proportions (or percentages), for instance, the proportion of people who declare that they are interested in politics. This is the `Politics` variable in the BSA. It has five categories ranging from 1 'A great deal' to 5- 'Not at all'. We could recode 1 and 2 - `quite a lot` into 'Significantly', but since we are only interested in estimating the confidence intervals, we will select the relevant values 'on the go'.

```
attr(bsa17$Politics,"label")     ### Phrasing of the question
```

```
## [1] "How much interest do you have in politics?"
```

```
attr(bsa17$Politics,"labels")     ### Value labels
```

```
## skip, version off route     Item not applicable       ... a great deal,
##                      -2                      -1                       1
##           quite a lot,                    some,          not very much,
##                       2                       3                       4
##        or, none at all?             Don`t know                 Refusal
##                       5                       8                       9
```

```
table(as_factor(bsa17$Politics)) ### Sample distribution
```

```
##
## skip, version off route     Item not applicable       ... a great deal,
##                       0                       0                     739
##           quite a lot,                    some,          not very much,
##                     982                    1179                     708
##        or, none at all?             Don`t know                 Refusal
##                     379                       1                       0
```

Note: Changes in a data frame are not automatically transferred into `svydesign` objects used for inferences. We therefore need to recreate it each time we create or recode a variable.

```
round(100*prop.table(svytable(~(Politics==1 | Politics==2),bsa17.s)),1)
```

```
## Politics == 1 | Politics == 2
## FALSE  TRUE
##    57    43
```

Let us now compute the confidence intervals for these proportions. Traditional statistical software compute these without giving us an idea of the underlying computations going on. Doing this in R

requires more coding, but also a better understanding of what is actually estimated.

Confidence intervals for proportions of categorical variables are usually computed as a sequence of binomial/dichotomic estimations – ie one for each category. In R this needs to be specified explicitly via the `svyciprop()` and `I()` functions. The former actually computes the proportion and its confidence interval (by default 95%), whereas the latter allows us to define the category we are focusing on.

```
svyciprop(~I(Politics==1 | Politics==2),bsa17.s)
```

```
##                                          2.5% 97.5%
## I(Politics == 1 | Politics == 2) 0.430 0.411  0.45
```

```
round(100*
        c(prop.table(svytable(~(Politics==1 | Politics==2),bsa17.s))[2],
attr(svyciprop(~I(Politics==1 | Politics==2),bsa17.s),"ci")),1
)
```

```
##  TRUE  2.5% 97.5%
##  43.0  41.1  44.9
```

Computing domain estimates

Computing domain estimates, that is estimates for subgroups adds a layer of complexity to the above example. They key point is that as weights were designed using the whole of the sample, computing estimates, in particular confidence intervals or standard errors for part of the sample, therefore using a fraction of these weights may affect the estimates. Instad it is recommended to use commands that take into account the entire distribution of the weights.

In R, the command that does this is `svyby()`

For instance, if we would like to compute the mean age of BSA respondents by Government Office Regions, we need to specify:

- The outcome variable whose estimate we want to compute: ie `RAgeE`
- The grouping variable(s) `GOR_ID`
- The estimate function we are going to use here: `svymean`, the same as we used before
- And the type of type of variance estimation we would like to see displayed ie standard errors or confidence interval

```
# bsa17$gor.f<-as_factor(bsa17$GOR_ID)
# bsa17.s<-svydesign(ids=~Spoint, strata=~StratID, weights=~WtFactor,data=bsa17)

round(svyby(~RAgeE,by=~as_factor(GOR_ID),svymean,design=bsa17.s,vartype = "ci")[-1],1)
```

```
##                             RAgeE ci_l ci_u
## A North East                 46.1 43.6 48.6
## B North West                 49.6 47.3 52.0
## D Yorkshire and The Humber   48.0 45.2 50.8
```

```
## E East Midlands              48.6 45.9 51.3
## F West Midlands              48.1 45.0 51.2
## G East of England            49.0 46.0 52.0
## H London                     45.0 43.0 46.9
## J South East                 48.0 45.1 50.8
## K South West                 53.4 51.5 55.2
## L Wales                      49.1 45.1 53.1
## M Scotland                   47.3 44.7 50.0
```

Note: we used `[-1]` from the object created by `svyby()` in order to remove a column with alphanumeric values (the region names), so that we could round the results without getting an error.

Our inference seem to suggest that the population in London is among the youngest in the country, and that those in the South West are among the oldest – their respective 95% confidence intervals do not overlap. We should not feel so confident about differences between London and the South East for example, as the CIs partially overlap.

We can follow a similar approach with proportions: we just need to specify the category of the variable we are interested in as an outcome, for instance respondents who are significantly interested in politics, and replace `svymean` by `svyciprop`.

```
round(
     100*
     svyby(~I(Politics==1 | Politics==2),
           by=~as_factor(GOR_ID),
           svyciprop,
           design=bsa17.s,
           vartype = "ci")[-1],
           1)
```

```
##                            I(Politics == 1 | Politics == 2) ci_l ci_u
## A North East                                            33.4 26.6 40.9
## B North West                                            41.9 36.1 48.0
## D Yorkshire and The Humber                              35.6 29.1 42.6
## E East Midlands                                         36.9 32.9 41.1
## F West Midlands                                         36.3 31.5 41.5
## G East of England                                       47.2 41.4 53.1
## H London                                                54.2 47.2 61.1
## J South East                                            44.6 38.7 50.8
## K South West                                            46.5 39.4 53.8
## L Wales                                                 38.6 27.7 50.7
## M Scotland                                              42.7 36.0 49.8
```

5.2 Inference with survey design variables using R

5.3 SDI Inference without design information using R

Example: count and proportion of the regional population of the UK using the LFS with End User License (EUL)

The EUL version of the LFS does not include sample design variables, just two weighting variables:

- `pwt22` for estimation with the whole sample
- `piwt22` for estimation using respondents currently in employment (typically used for earnings estimation)

```
svyset [pw=pwt22]
svy:tab uresmc, cell count percent format(%10.1g)
```

Syntax Using R

```
library(survey)
lfs.s←svydesign(ids=~1,weights=~pwt22,data=lfs) ### Assuming the dataset is stored as lf
svytable(lfs.s)
```

5.4 SDI Inference without design information using R

5.5 Point estimates using casual weighting

See here for the Labour for Survey and for the family resource survey - For the FRS: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/972808/Ch1_Methodology_and_Standard_Errors.xlsx - For the LFS: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesno9guidetocalculatingstandarderrorsforonssocialsurveys#annex-a-labour-force-survey-standard-errors-january-to-march-2015-united-kingdom

- use point estimates with their standard errors using the survey design commands available in most statistical software. These will assume that the data was collected under simple random sampling using provided weights then adjust them using the design factors published by the data providers (LFS). In some cases,

Example: count and proportion of the regional population of the UK using the LFS

Syntax Using Stata

Secure access data Produce estimates using the sample design variables (ie clusters and/or strata) and the specialist survey functions of existing statistical software.

Estimating quantities about subgroups in the data

Using EUL data In the case of the LFS, standard errors would likely to be overestimated, and therefore the estimates would be conservative, whereas in the case of the FRS, these would be underestimated. The seriousness of these would increase with smaller subgroups, therefore users should try and avoid working with groups that are too small. When estimating quantities for

domain, it is also recommended to use functions that explicitly take into the grouping rather than only working with the subpopulation of interest. Not doing so could lead to incorrect weighting of estimates. In case of simple domain estimates, it might still be possible to rely on estimates published by the data producer.

In the case of the LFS

In the case of the FRS

Using secure lab data Point and reliability estimates may be computed using the survey-based estimation commands.

Sample design and multivariate analysis ie regression

TBC – same message as previous section+ issue of controlling for vs using weights.

6. SPSS Examples

Standard editions of SPSS do not include support for survey design variables, and only limited use of sampling weights. When using grossing weights – ie weight that have been designed to enable computing population totals from sample data – as is the case for instance with the Labour Force and Family Resources surveys, measures of dispersion and standard errors will not be adequately computer. It is therefore not recommended to attempt using the base version of SPSS with survey data beyond estimating point estimates. Significance test, and standard errors will not reflect the correct values. USers willing to use SPSS with survey data will need to acquire the Premium Edition or the Complex Samples option of the software.

(TBC)

7 Stata examples

Stata provides comprehensive support for computing estimates from survey data. Users may either opt to add sampling weights to the standard estimation commands, or use survey-specific commands. The latter is recommended when knowledge of estimate precision is required. Stata provides a conceptual distinction between four types of weights: Frequency weights, Variance weights, Importance weights and Probability weights. These differences impact on the way standard errors are computed. In most cases, social survey weights from UKDS datasets should be treated as probability weights. A number of of basic estimation commands, such as *summarise* do not allow using probability weights. This is an explicit features of Stata, meant to nudge users of survey data to prioritise the survey commands rather than 'casual' weighting.

Using standalone weight specification (ie not using survey design functions). In Stata it consists in the weighting variable being specified between square brackets. Stata defines four kind of weights: frequency weights (`fweight`), analytical weights (`aweight`), importance weights (`iweight`) and probability weights (`pweight`). Technically speaking, only the latter (abbreviated as `pw` in most Stata commands) should be used with survey data. However, Stata does not allow using probability weights standalone with its main commands, for the reason highlighted above ie in order for users

not overlook survey design issues in their data. Therefore, one has to specify instead the wrong frequency weights (`fw`) if one does not wish to use the survey design functions.

(TBC)

8. References & further information

UKDS (2019) Weights in social surveys: an introduction: https://www.youtube.com/watch?v=Vllr4olp3N4&t=39s

Goldsmiths(2020) W7: Using survey weights in R https://www.youtube.com/watch?v=brxx81U6N1o

Datacamp (2020) R Tutorial: What are survey weights? https://www.youtube.com/watch?v=8iMV7ei61IM Note: basic, partially available, complex survey design in R

DWP (2014) Uncertainty in Family Resources Survey-based analysis. Guidance on estimating uncertainty in Family Resources Survey-based analysis. https://www.gov.uk/government/publications/uncertainty-in-family-resources-survey-based-analysis

UKDS (2018) Data Skills Modules: Applying weights to survey data https://www.youtube.com/watch?v=TIad5__WP8g Note: point and click howto in SPSS

Curran (2016) Complex Survey Designs and Weighting Using Stata: Part 1-3, https://www.youtube.com/watch?v=oOpJdC_oeKY

ONS (2022) Family Resources Survey, 2020/21 Methodology and Standard Error Tables https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1065513/Ch1_Methodology_and_Standard_Errors.ods

https://www.ibm.com/support/pages/inconsistency-output-when-using-weighting-procedure