Using weights to produce population inference in UKDS surveys – a practical guide

2022-12-23

# Introduction

This note aims at signposting users of UKDS social surveys towards the appropriate usage of weights and survey design variables when conducting inferential analysis as well as providing guidelines in order to safely do so. This preliminary version of the document only provides examples from the Labour Force and Family Expenditure Surveys but will be gradually augmented over time.

Weights in social surveys are designed to compensate for the under- or over-representation of groups at the time the sample was drawn. How to use them in analyses of social surveys depends on a wide range of factors including the scope of the analysis conducted, the chosen statistical methods, the license version of the dataset, the capabilities of the statistical software, as well as the sample design and the weights provided. Using weights and more generally, accounting for the sample design raise complex issues for which in most cases no ideal solution exist.

The present document focuses on a limited number of practical procedures to achieve the above. It does not provide an in-depth discussion of the theoretical underpinning, survey design or sampling of the data.

The content of this note summarises technical information provided by the Office for National Statistics. (Ref to be added)

# Whole population estimates (not subgroups)

## Simple point estimates (ie mean, median, proportion, total) of a variable.

Example: the mean/median age, the proportion of people in employment, the total number of men and women.

### Cases where knowing about the precision of estimates is not required

These are cases where users are not interested in obtaining the standard errors of point estimates, their confidence interval or conduct statistical testing, for example because they are simply learning or teaching basic statistical concepts or how to use software. They are nonetheless interested in having an idea of the value of a parameter in the population. Users can either:

* Compute the weighted estimates using available commands in mainstream statistical software (R, SAS, SPSS, Stata) that accepts weights. Most of these will provide the correct value of the estimates.
* Users may also use the survey specific estimation commands of these packages. This is not strictly necessary, but is recommended as it will help users remain aware that inference with survey data requires special attention.

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

**Syntax Using Stata\* ie mean yes, median no**

tab uresmc [fw=pwt22]

**Syntax Using R**

xtabs(pwt22~uresmc,lfs)

**Syntax Using SPSS**

### Cases where knowing about the precision of estimates is required

This happens when users need to know about the standard error of an estimate, its confidence interval or compute a statistical test, for example because the data will be used in a scientific article, or will be published.

**EUL datasets (without information about sample design)** End User Licence datasets released by UKDS are not allowed to include potentially disclosive information, which unfortunately includes information about clustering, small geographical areas or primary sampling unit. It is therefore not possible to directly compute robust information about estimate precision using these.

Users can either: - Directly use published estimates and standard errors already computed by the data producer for some of their social surveys; - Adjust the standard error of their estimates using Design factors published by data producers. A design factor is a number by which to multiply standard errors estimated under the assumption of simple random sampling.

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*

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 lfs  
svytable(lfs.s)

*Syntax Using SPSS* TBC

**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 This covers estimates of quantities about subgroups of the populations (aka domains) rather than the whole population (ie groups whose size is itself an estimate). Example: mean age by gender, or some other categories, or analysis restricted to a subset of the population (for example those in employment).

**Using EUL data** There is no direct way of estimating unbiased domain estimates with the information provided in EUL datasets. The closest users can come to it to use survey procedures available in statistical software assuming random sampling, report the results as confidence intervals rather than point estimates in order to convey the uncertainty of the result, and document the likely effect of not taking into account the survey design ie over or underestimation of the actual standard error. 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](https://www.nisra.gov.uk/publications/labour-force-survey-annual-report-2021)

[In the case of the FRS](https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesno9guidetocalculatingstandarderrorsforonssocialsurveys#annex-a-labour-force-survey-standard-errors-january-to-march-2015-united-kingdom)

**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.

# Study-specific Weight and sample design information

## 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)

## Family Resources Survey

The LFS 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

# Software notes

**R** Being open source, R does not provide a centralised/unified sets of command to compute weighted estimates and accounting for sample design. The algorithms implementations of statistical theory may vary between packages, but are usually provided in the package documentations. In practice, computing weighted point estimates is straightforward. The *Hmisc* package provides a number of useful functions allowing to do so. The *Survey* package provides a comprehensive set of function for computing point estimates and reliability from survey data. It is recommended to use this for other usages than casual weighting.

**SAS** TBC

**SPSS** 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.

**Stata** 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.

# 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>