European
Social
Survey

ESS Sample Design Data Files

The ESS Sampling Expert Panel

Stefan Zins

29th June 2016

http://www.europeansocialsurvey.org/

Summary

This document reports on the creation and use of the ESS Sample Design Data File (SDDF). The SDDF is routinely generated by the National Coordinator after fieldwork has finished. It includes information on the implemented sample design such as inclusion probabilities and clustering. As such, it serves the sampling team with the data required for computation of design weights, design effects and as a general basis for benchmarking the quality of sampling. The ESS user may use it for several purposes such as incorporating cluster information in her analyses. This documentation aims at clarifying important issues connected with the creation and the use of the SDDF.

Contents

1 Included Variables

In the SDDF, information is given on a set of seven variables for every country and every ESS round. These variables are `cntry`, `idno`, `psu`, `domain`, and `prob`. They are described in detail in the following section.

1.1 CNTRY

The two-letter code country abbreviation string variable identifies different countries. When merging SDDF data to the integrated file (i.e. multiple countries) using `idno`, `cntry` must be used in combination with `idno` to avoid ambiguous matches on `idno` since there might exist identical `idno`s in different countries.

1.2 IDNO

The individual identification number serves as a unique sample person identifier within a given country. It can be used to merge sample data and the SDDF from the same country (see above).

1.3 PSU

This variable includes information on the primary sampling unit (PSU). Respondents belonging to the same primary sampling unit will have the same value on PSU. This variable is mainly useful when design effects are to be considered.

1.4 DOMAIN

This variable indicates, if applicable, to which sampling domain a unit belongs within a country. Some countries deploy different sampling design independently within separat areas, also called sampling domains. For instance, a single-stage design, where individuals are selected directly, might be used in densely populated metropolitan areas and a multi-stage design for the rural areas of a country. Because of cost reasons it is often not feasible to have the sampling unit scattered across a wide area. Thus, for less densely populated areas settlements or municipalities have to sampled first as PSUs and then persons are selected within PSUs. Hence, there are two sampling domains, the metropolitan areas and the rural areas.

1.5 STRATIFY

The identifier of stratum membership of the unit, if stratification was used at the first sampling stage. If the sampling of the PSUs included stratification the `stratify` indicates to which stratum the PSUs and thus the respondents belong.

1.6 PROB

The probability for the respondent of being selected into the gross sample, also called inclusion probability. It is the basis for the design weights of the ESS.

2 Using the SDDF

SDDF data can be used to enrich and improve your analyses. The most common use will be to generate weights as well as including PSU information in a multi-level model or to estimate design effects for specific variables. The following sections explain each of these uses.

2.1 Using Inclusion Probabilities to compute Weights

For convenience of illustration, values of PROB shall be denoted by π_i, where $i = 1, \ldots, n$ and n is the sample size. The inverse of π_i is simply the raw design weight and is formally defined as

$$w_i = \frac{1}{\pi_i}$$

An important feature the raw weights is that their sum $\sum_i^n w_i = \hat{N}$ is an unbiased estimator for N, the target population size. The raw weights are very huge numbers and one might want to rescale them to a more convenient range. One possibility is to scale the raw design weights to the net sample size n_{net}. This is done by the following simple transformation:

$$\tilde{w}_i = n \times \frac{w_i}{\sum_i^n w_i} \ .$$

Now the scaled weighs $\tilde{w}_i$ have the property that thier sum $\sum_i^n \tilde{w}_i = \mathrm{n}_{net}$ and consequently their mean is equal to one.

Example:
We can see the difference between a weighted and an unweighed estimate in the following example. These differences can become especially apparent if we are estimating statistics that are sensitive to the tails of the sampling distribution of our variables of interest. For instance, assume that we are interested in the estimating the 10% quantile for varaible `ppltrst`, $q_{10}(ppltrst)$, of the Netherlands based on the ESS round 7 data. The weigthed estimator has a value of $\hat{q}_{10\,w}(ppltrst) = 5$ where the unweighted estimator gives $\hat{q}_{10}(ppltrst) = 4$.

3 Using PSU Identifiers for the Estimation of Design Effects

Design effects arise from a variety of divergences in real-world sample surveys from the ideal of simple random sampling (SRS). Most prominent and intuitively appealing is the *design effect due to clustering*, abbreviated in the following by Deff_c. Due to the fact that respondents living in the same geographical area are socialised in similar ways, their responses to survey questions resemble each other more than they resemble the responses of another geographical area. However, the fact that the responses are more similar implies that a sampling design that selectes such clusters, say at the first sampling stage, can produce less acurate estimates than a SRS of the same size. This in turn means that the variance and also the standard error of an estimator, $\hat{\theta}$, is (usually though not necessarily) underestimated by the variances estimator usually used for SRS. The factor by which the variance is underestimated is the design effect.

The most basic and thus best known definition of the design effect is given in Kish (1965) where Deff_c is defined as (see also Ganninger, 2010)

$$\text{Deff}_c = \frac{\text{V}_{clu}(\hat{\theta})}{\text{V}_{SRS}(\hat{\theta})}$$

where $\text{V}_{clu}(\hat{\theta})$ is the variance of the estimator $\hat{\theta}$ under the actual cluster design and $\text{V}_{SRS}(\hat{\theta})$ is the variance of the same estimator under a (hypothetical) simple random sample. Kish (1965) described this quantity in the following way

$$\text{Deff}_c = 1 + (\bar{b} - 1)\rho \ ,$$

where $\bar{b}$ is the average cluster size and ρ the intraclass correlation coefficient for the cluster units, i.e. the correlation between responses within the same cluster. Gabler et al. (1999) and Gabler et al. (2006), showed that there exists a model-based justification for the above formula for

$$\hat{\theta} = \overline{y}_w = \frac{\sum_i^n y_i \tilde{w}_i}{\sum_i^n \tilde{w}_i} \ ,$$

which yields a model-based design effect which is the product of the design effect due to unequal selection probabilities (Deff_p) and Deff_c and is defined as

$$\text{Deff} = \text{Deff}_p \times \text{Deff}_c = n \frac{\sum_i^n \tilde{w}_i^2}{(\sum_i^n \tilde{w}_i)^2} \times [1 + (b^* - 1)\rho]$$

where

$$b^* = \frac{\sum_c^C (\sum_i^{n_c} \tilde{w}_{ci})^2}{\sum_i^n \tilde{w}_i^2}$$

with $c = 1, \ldots, C$ is an index for the clusters or PSUs and $j = 1, \ldots, n_c$ denotes elements within the c-th cluster or PSUs. To estimate the design one needs to estimate the ρ. One way to do this is to use the an ANOVA estimator $\hat{\rho}_{avo}$, which can be wirten as

$$\hat{\rho}_{avo} = \frac{MSB - MSW}{MSB + (K - 1)MSW} \ ,$$

with

$$MSB = \frac{SSB}{C-1} \ ,$$

where $SSB = \sum_c^C n_c(\overline{y}_c - \overline{y})^2$ and

$$MSW = \frac{SSW}{n-C} \ ,$$

with $SSW = \sum_c^C \sum_i^{n_c} (y_{ci} - \overline{y}_c)^2$ and

$$K = (C-1)^{-1} \left(n - \sum_c^C \frac{n_c^2}{n} \right) \ ,$$

where $\overline{y}_c = \sum_i^{n_c} y_{ci} {n_c}^{-1}$ the mean of the c-th cluster and $\overline{y} = \sum_i^n y_i n^{-1}$ as the sample mean.

For sampling design that do not use cluster sampling or select only on element in each cluster $b^* = 1$ (also $\rho = 0$), thus we have $\text{Deff}_c = 1$, i.e. there is no cluster effect. If the survey weights are equal for all elements in the sample $\text{Deff}_c = 1$, else $\text{Deff}_c > 1$, which means that unequal survey weights will increase the estimator of the design effect.

Example:
We want to estimate the design effect for $\overline{y}_w(ppltrst)$, the weighted mean of variable `ppltrst`, for Austria in ESS round 7. Thus we calculate the following values, using the `psu` variable from the SDDF and dweight variable from the Austrian data file (by merging both data sets with the help of the variable `idno`):

$$\begin{aligned} \text{Deff}_p &= 1.0472956 \\ b^* &= 5.3698456 \\ MSB &= 30.3307619 \\ MSW &= 5.2384671 \\ K &= 5.1411104 \\ \hat{\rho}_{avo} &= 0.4823231 \\ \text{Deff}_c &= 3.1076774 \end{aligned}$$

Finally we have $\text{Deff} = \text{Deff}_c \times \text{Deff}_p = 3.2546568$.

References

Gabler, S., **Häder, S.** and **Lahiri, P.** (**1999**): *A model based justification of Kish's formula for design effects for weighting and clustering.* Survey Methodology, 25 (1), pp. 105–106.

Gabler, S., **Häder, S.** and **Lynn, P.** (**2006**): *Design Effects for Multiple Design Samples.* Survey Methodology, 32 (1), pp. 151–120.

Ganninger, M. (**2010**): Design Effects: Model-based versus Design-based Approach. Mannheim: GESIS - Leibniz-Institut fÃ¼r Sozialwissenschaften.

Kish, L. (**1965**): Survey Sampling. New York: John Wiley & Sons.