The ESS Sample Design Data File (SDDF)

Documentation

Version 1.0

Matthias Ganninger

Tel: +49 (0)621 1246 – 282

E-Mail: matthias.ganninger@gesis.org

April 8, 2008

Summary: This document reports on the creation and use of the ESS Sample Design Data File (SDDF). The SDDF is routinely generated by an ESS country's 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 analyst 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.

1 Included Variables

In the SDDF, information is given on a set of six variables for every country and every ESS round[1]. These variables are CNTRY, IDNO, STRATIFY, PSU, SAMPPOIN, and PROB. They are described in detail in the following section.

1.1 CNTRY

The two-letter code country abbreviation string variable identifies different ESS countries. When merging SDDF data to the integrated file using IDNO, CNTRY must be used in combination with IDNO to avoid ambiguous matches on IDNO since there might exist identical IDNOs 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 STRATIFY

This variable is a marker for the combination of all stratification variables implemented in a certain country. If there is, for example, explicit stratification by regions as well as implicit stratification by systematic sampling of addresses, STRATIFY combines the labels of each of the two single stratification variables into one string (e.g. "23-42" indicating stratum 23 on the explicit stratification variable and stratum 42 on the implicit stratification variable).

1.4 PSU

This variable includes information on the primary sampling unit (i.e. *cluster*). Respondents belonging to the same primary sampling unit will have the same value on PSU. This variable is mainly useful when considering the design effect due to clustering.

1.5 SAMPPOIN

In some countries, PSUs are not the ultimate clusters. In these cases, a lower-level structure, called sampling point, exists. A good example for this separation is the ESS round II sample data in Portugal. Here, a total of 100 localities (PSUs) was surveyed.

[1] The user may also wish to consult the sampling plans provided for each country at http://ess.nsd.uib.no

However, within each locality, a certain number of starting addresses was sub-sampled (SAMPPOIN) which form the ultimate clusters.

1.6 PROB

The product of a respondent's inclusion probability on each stage is captured by PROB. It can thus serve as a basis for weighting issues.

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 or for variance estimation. The following sections explains some of these uses.

2.1 Using Inclusion Probabilities to compute Weights

As mentioned above, PROB stores the product of a sample element's inclusion probabilities on every stage. 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

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

Example:
In this example, we see how inclusion probabilities of different stages transform to PROB and how this overall inclusion probability is transformed to diverse weights. In this and all following examples, ESS round II data from France are used for illustration. The sample design can be summarized as follows: On the first stage, 200 primary sampling units (communities), are sampled.
A PSU has a given inclusion probability, denoted by PROB1. Then, on the second stage, households are selected. Each household also is associated a certain probability of inclusion, PROB2. On the third stage, a respondent within a selected household is sampled via last-birthday method. His or her probability of inclusion is simply the inverse of the number of persons belonging to the target population. Our study variable shall be the overall satisfaction with life (STFLIFE). A respondent may have the following characteristics:

No.	IDNO	PROB1	PROB2	PROB3	STFLIFE
18	102010	.004731855	.03612479	0.25	6

The product of the three inclusion probabilities is $.004731855 \times .03612479 \times .25 = .00004273432$. Taking the inverse of this number, we end up with a raw weight of $w_{18} = 23400.4$, which equals the number of population units this respondent represents.

The raw weights are very huge numbers and one might want to rescale them to a more convenient range. One possibility is to normalize the raw design weights to the net sample size. This is done by the following simple transformation:

$$\tilde{w}_i = n \times \frac{w_i}{\sum_{i=1}^{n} w_i}. \tag{2}$$

Finally, in extreme cases, weights greater than 4.0 are truncated to this threshold. Usually, this is necessary only in very few countries and very few cases.

Example:
We can see the difference between a weighted and an unweighed estimate in the following example. Le us return to the above case and assume we computed raw weights for all sample elements. Assume we are interested in the usual Horvitz-Thompson estimator of the sample mean which is defined by

$$\bar{y}_{HT} = \frac{1}{N} \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \frac{1}{N} \sum_{i=1}^{n} y_i \times w_i$$

That means, we simply multiply STFLIFE with the corresponding weight, take the sum over all sample elements and divide it by the sum of weights. If we do so with our sample data set, we get 6.44 as an estimate of the average satisfaction with life. The unweighted sample mean, $\frac{1}{n} \sum_{i=1}^{n} y_i$, however, is 6.37 and thus deviates from the weighted one.

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. 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 (PSU or sample point) 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, in terms of precision,

the cluster sample data correspond to simple random sample data with less responses. This in turn means that the variance and also the *standard error* of an estimator, $\hat{\theta}$, is underestimated by the naive formula. 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

$$deff_c = \frac{Var_{clu}(\hat{\theta})}{Var_{srs}(\hat{\theta})}, \tag{3}$$

where $Var_{clu}(\hat{\theta})$ is the variance of the estimator $\hat{\theta}$ under the actual cluster design and $Var_{srs}(\hat{\theta})$ is the variance of the same estimator under a (hypothetical) simple random sample. Kish (1965) also showed that this quantity can be expressed as

$$deff_c = 1 + (\bar{b} - 1)\rho, \tag{4}$$

where $\bar{b}$ is the average cluster size and ρ is the intra-class correlation coefficient. Gabler et al. (1999) and Gabler et al. (2006) showed that there exists a model-based justification for the above formula which yields a model-based design effect. It is the product of the design effect due to unequal selection probabilities ($deff_p$) and $deff_c$ and is defined as

$$deff = deff_p \times deff_c = n\frac{\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2} \times \left[1 + (b^* - 1)\rho\right], \tag{5}$$

where ρ is the intraclass correlation coefficient and

$$b^* = \frac{\sum_{c=1}^{C}\left(\sum_{j=1}^{b_c} w_{cj}\right)^2}{\sum_{i=1}^{n} w_i^2}, \tag{6}$$

where $c = 1, \ldots, C$ is an index for clusters and $j = 1, \ldots, b_c$ denotes elements within a given cluster c of size b_c.

Note that it makes no difference which type of design weights are used, normalized or raw weights. However, as a rule, we use normalized weights, so we set $w_i = \tilde{w}_i$ and $w_{cj} = \tilde{w}_{cj}$ in (5) and (6).

The information on inclusion probabilities and on PSU given in the SDDF enables the user to estimate the design effect due to unequal selection probabilities as well as the design effect due to clustering.

Example:
Returning to the ESS round II data of France, according to (5) $\widehat{deff}_p$ is computed in the following way:

1. compute the sum of squared weights: $\sum_{i=1}^{n} \tilde{w}_i^2 = 2157.239$,
2. compute the squared sum of weights: $(\sum_{i=1}^{n} \tilde{w}_i)^2 = 3261636$,
3. insert the values into the formula: $\widehat{deff}_p = 1806 \times \frac{2157.239}{3261636} \approx 1.19$.

For estimation of $deff_c$ for a specific variable, we first have to reduce the dataset to only those cases where the variable under study is not missing.

Example:
Let us again take the ESS II data of France. We want to estimate $\widehat{deff}_c$ for STFILFE. We just need to

1. compute $b^* = \frac{19604}{2155} = 9.09$,
2. calculate (e.g. the ANOVA) estimator of ρ, which is $\rho_{ANOVA} = 0.0373$,
3. and then insert these values into the formula: $\widehat{deff}_c = 1 + (9.09 - 1) \times 0.0373 \approx 1.3$

The design effect of STFLIFE is simply the product of $\widehat{deff}_p$ and $\widehat{deff}_c$ as in (5).

Example:
The design effect for the French ESS II STFLIFE variable then is the product of $\widehat{deff}_p$ and $\widehat{deff}_c = 1.19 \times 1.3 = 1.547$.

This means that the true variance of the sample mean for STFLIFE is 1.547 times larger than the naive formula $\mathrm{Var}(\bar{x}) = \frac{\mathrm{Var}(x)}{n}$ would suggest. It thus also implies that the standard error is 1.24 ($\sqrt{\widehat{deff}}$) times as large as the one estimated under the assumption that the data come from a simple random sample.

References

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

Gabler, S., Häder, S. & Lynn, P. (2006), 'Design effects for multiple design surveys', *Survey Methodology* **32**(1), 115–120.

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