Documentation of the Dataset DESIGN of the Socio-Economic Panel Study (SOEP)

Martin Spiess
mspiess@diw.de

Martin Kroh
mkroh@diw.de

May 30, 2007

1 Sample Design of the SOEP

Information on the sample design of the SOEP that was previously compiled in the dataset VARIANZ (Spiess 2001) will be disseminated in a revised and amended dataset DESIGN from the 2007 data distribution on. This preliminary short documentation provides an overview of the variables PSU, STRAT, and DESIGN included in the dataset. Additional information on the random group variable RGROUP and the identifier of the interviewer conducting the first-wave interview INTNR can be found in the general introduction to the SOEP (Haisken-DeNew & Frick 2005).

Depending on the specific requirements and scope of the survey at the time of introducing each of its subsamples, SOEP-samples A through H were drawn using somewhat different sampling schemes. While, for instance, sample C from 1990 is a two-stage clustered but unstratified sample of East German private households, sample G from 2002 is an unclustered sample of high-income households stratifying for East and West Germany and for income. The dataset DESIGN provides information necessary to replicate the specific sampling designs of its subsamples. For a general discussion of sampling methods and strategies for adjusting applied analyses to these designs, see, for instance, Särndal, Swensson, & Wretman (1993).

2 Primary Sampling Units

As is the case in most general population surveys, six out of eight SOEP subsamples were drawn in a first stage from small-scale regional units (primary sampling units) and in a second stage from households within these regional units (secondary sampling units). The regional units of the first sampling stage correspond largely to the electoral districts for the German National Assembly. The variable `PSU` uniquely identifies the primary sampling units from which households were drawn in each subsample of the SOEP. Note that due to data protection laws, we assigned the identifying numbers randomly. The number of primary sampling units varies by samples. While samples D and G are unclustered, i.e., composed of only a single `PSU`, the number of primary sampling units ranges from 40 in sample B to 985 in sample F.

3 Stratification and Design Weights

We conceive of SOEP's basic structure, which consists of eight subsamples (A through H), as a form of stratification in the overall SOEP sample. Due to the differing sizes and in part differing underlying populations (West Germans, Immigrants, East Germans, etc.) of each subsample, the households drawn in each have different inclusion probabilities. Moreover, individual subsamples may comprise several strata with–again–varying inclusion probabilities. Sample B is stratified according to the household size and the nationality of the head of household (Turkish, former Yugoslavian, Italian, Greek, and Spanish). Sample D stratifies for migrants from the former Soviet Union, migrants from the former GDR, and other migrants coming to West Germany between 1984 and 1994. Sample F oversamples households with migration background relative to the rest of the population, i.e., composed of two strata with different inclusion probabilities. Finally, sample G is stratified for households from East and West Germany and for households with a monthly income between 7,500 DM and 10,000 DM and households with an income greater than 10,000 DM. All other SOEP subsamples (A, C, E, and H) are unstratified and each therefore has a constant inclusion probability. Spiess (2005) discusses the derivation of these inclusion probabilities for subsamples A through F. We are currently working on an integrated and updated documentation of the SOEP sampling design that describes all sampling information on SOEP filed in the dataset `DESIGN` and the derivation of these information for all subsamples, including the most recent samples G and H.

The dataset DESIGN provides information on the stratified sampling of the SOEP in form of two variables. The variable STRAT identifies each of the discrete sampling groups described above. Altogether, the SOEP consists of 40 strata: one stratum in sample A, twenty-seven in sample B, one in sample C, three in sample D, one in sample E, two in sample F, four in sample G, and one in sample H. Unique inclusion probabilities pertain to each of these strata. The variable DESIGN contains the inverse of this probability, i.e., the design weight (Horvitz & Thompson 1952).

4 References

Haisken-DeNew, J.P. and Frick, J.R. (eds.). 2005. *Desktop Companion to the German Socio-Economic Panel (SOEP)*. Berlin: DIW Berlin.

Horvitz. D.G. and Thompson, D.J. 1952. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association* 47: 663-685.

Särndal, C.-E., Swensson, B. and Wretman, J. 1993. *Model Assisted Survey Sampling*. New York: Springer.

Spiess, M. 2001. *Description of the Variables: STRAT1, STRAT2, and SAMPOINT*. Berlin: DIW Berlin.

Spiess, M. 2005. *Derivation of Design Weights: The Case of the German Socio-Economic Panel (SOEP)*. DIW Data Documentation 8.