DIW **Materialien**
Research Notes

Research Note No. 5

Derivation of design weights: The case of the German Socio-Economic Panel (GSOEP)

by
Martin Spiess

Berlin, October 2001

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 14195 Berlin
Phone: +49-30-89789- 0
Fax: +49-30-89789- 200
Internet: http://www.diw.de

Derivation of design weights: The case of the German Socio-Economic Panel (GSOEP)

(Updated and revised version of Discussion Paper No. 197)

Martin Spiess[a]

Abstract. Design-based estimators of totals, means or proportions in finite populations generally are functions of weighted sums. If each element selected into the sample is also observed, then for the calculation of the π-estimator these weights are just the inverse inclusion probabilities of the elements. However, if e.g. nonresponse or attrition over time occurs, calculation of these weights also includes modeling of nonresponse and/or attrition mechanisms. Since models of these mechanisms are disputable, 'pure' design weights can be the basis for calculating alternative weights by a different modeling e.g. of nonresponse and/or attrition mechanisms. In the case of complex sampling schemes, however, it is often not possible to derive the exact inclusion probabilities. In that case, weights may be derived based on approximations and/or simplifying assumptions. In this paper, after describing the selection schemes of the subsamples A, B, C, D, E and F of the German Socio-Economic Panel (GSOEP), approximate design weights are derived which enable users of the GSOEP to calculate their own weights if desired.

Key words: Design-based inference; approximate design weights; complex surveys; GSOEP

[1]Longitudinal Data and Microanalysis, DIW Berlin, Königin-Luise-Str.5, 14195 Berlin, Germany

1 Introduction

Design-based estimators of totals, means or proportions in finite populations generally are functions of weighted sums. If each element selected into the sample is also observed, then for the calculation of the π-estimator these weights are just the inverse inclusion probabilities of the elements. However, if e.g. nonresponse or attrition over time occurs, calculation of these weights also includes modeling of nonresponse and/or attrition mechanisms. Since models of these mechanisms are disputable, 'pure' design weights can be the basis for calculating alternative weights by a different modeling, e.g. of nonresponse and/or attrition mechanisms. In the case of complex sampling schemes, however, it is often not possible to derive the exact inclusion probabilities. In that case, weights may be derived based on approximations and/or simplifying assumptions.

In this paper, the calculation of approximate design weights, i.e. approximate inverse inclusion probabilities of households at the first wave of each subsample of the German Socio-Economic Panel (GSOEP), will be described. In deriving these weights, nonresponse is not accounted for. Therefore, if corresponding estimators are to be calculated, then nonresponse mechanisms or (if not only first-wave samples but also those of later waves are used) e.g. attrition mechanisms have to be modeled separately. The weights derived are given in Table 1 in the Appendix. They are stored under the name DESIGN in file VARIANZ on the GSOEP CD.

The GSOEP consists of several subsamples (denoted as subsample A–F) starting at different points in time (Wagner et al., 1994). Description of the sample inclusion schemes can be found in Pannenberg et al. (1998) or Rendtel (1995). In the following sections, for each subsample the calculation of the approximate weights are given after a short description of the sample schemes. It must be noted, however, that the schemes used to select the different samples are complex, and not all information needed to calculate exact inclusion probabilities are available. Therefore, the formulas used to derive the weights have to be interpreted as models of the underlying process, thus switching from an approach traditionally denoted as design-based to a model-based approach. The derivation of the 'true' inclusion probabilities of households is very similar for subsamples A–C, E and F. Therefore, only for subsample A, selected in 1984, the main problems of the derivation of 'true' inclusion probabilities of households will be

discussed. Concerning a discussion of the selection scheme used for subsample D, the interested reader is refered to Rendtel, Pannenberg and Daschke (1997).

2 Basics and Notation

In this section, the notation is introduced and some basic results are given. For details concerning the concepts used in this section, see e.g. Särndal, Swensson and Wretman (1993).

Let s be a specific sample, regarded as the outcome of a set-valued random variable S, and $Pr(S = s) = p(s)$ be the probability of selecting s under a given sample selection scheme. The function $p(s)$ or p for short, is called the sampling design. The ingredients needed for determining the inclusion probability π_k of an element k given the basic systematic sampling scheme, are the fixed sampling interval, denoted as a, the number of population elements N, and n, the integer part of N/a (see, e.g. Särndal, Swensson and Wretman, 1993, pp. 73–75). Then $N = na + c$, where $0 \leq c < a$. If $c = 0$, the sample size is n. If $c > 0$, the sample size is either n or $n + 1$. The probability for every possible sample given this design is $p(s) = 1/a$, and the probability of selecting element k is given by $\pi_k = 1/a$. Given the circular systematic sampling method, the inclusion probability is $\pi_k = n/N$ where n is the sample size (Särndal, Swensson and Wretman, 1993, p. 77). Now, turning to a systematic probability proportional-to-size without replacement scheme (systematic πps scheme; Särndal, Swensson and Wretman, 1993, p. 96), let x_k be a positive and known auxiliary variable or size measure of element k. Furthermore, let $T_0 = 0$, $T_k = T_{k-1} + x_k$ $(k = 1, \ldots, N)$, a be a fixed integer (the sampling interval) and n again be the integer part of T_N/a, where $T_N = \sum_U x_k$ and $\sum_U$ means summation over all population elements. Then $T_N = na + c$, where $0 \leq c < a$. If $c = 0$, the sample size is n. If $c > 0$, the sample size is either n or $n + 1$. Assume that $nx_k \leq T_N - c = na$ for all k and, for simplicity, assume that x_k is an integer. The probability of a specific sample is $p(s) = 1/a$, and the inclusion probability π_k is given by

$$\pi_k = \frac{nx_k}{T_N - c} = x_k \frac{1}{a}.$$

If a sample is selected according to a selection scheme with more than one phase or stage (for the differences between two-phase or multiphase and two-stage or

multistage designs, see Särndal, Swensson and Wretman, 1993, p. 133 ff and p. 343 ff), then for the first phase or stage the index 1 will be used and for the second phase or stage the index 2 will be used. For example, a first-phase sample will be denoted as s_1, its sampling design as $p_1(s_1)$ and so on. Because of redundancy, no index will be used for the last phase or stage sample.

3 Subsample A

The population from which subsample A was selected was defined to be the set of private households where the household head did not have the Turkish, Italian, Greek, (former) Yugoslavian or Spanish nationality. Subsample A was selected in 1983/1984. The sampling scheme has two stages and two phases within the first stage. In the first phase of the first stage, the primary sampling units (units smaller than constituencies, i.e. 'Stimmbezirke'; PSUs) were selected, and in the second stage the secondary sampling units (households; SSUs) were selected. The scheme used to select the first-phase-first-stage sample of PSUs may be described as a systematic probability proportional-to-size without replacement scheme (systematic πps–scheme, see, e.g. Särndal, Swensson, and Wretman, 1993, p. 96). However, since the sizes of the PSUs, given by the number of households belonging to the defined population, were unknown, they had to be estimated. This first-phase-first-stage sample was then stratified so as to mimic certain marginal distributions according to several variables very similar to those variables used to sort the population elements (PSUs) to select the first-phase-first-stage sample. Within each cell, again samples of PSUs according to a systematic πps–scheme were selected. Given this second-phase-first-stage sample of PSUs, within each PSU, the SSUs (households) were selected according to a scheme that may approximately be described by a circular systematic sampling scheme with random start (see, e.g., Särndal, Swensson, and Wretman, 1993, p. 77).

Given the above sampling scheme and assuming that all necessary quantities are known, the inclusion probability of household h in PSU k is given by

$$\pi_{h,k} = \sum_{s_1 \ni k} \sum_{s_2 \ni k} \sum_{s \ni h} p(s|s_1, s_2) p_2(s_2|s_1) p_1(s_1),$$

where $s_1 \ni k$, $s_2 \ni k$ and $s \ni h$ means taking the sum over those samples s_1, s_2 and s that contain the given kth PSU or hth household, respectively. Following

the model of a circular systematic sampling scheme for the second-stage sample, the inclusion probability of household h given the first- and second-phase-first-stage samples s_1 and s_2, is given by $\pi_{h,k|s_1,s_2} = \sum_{s \ni h} p(s|s_1, s_2) = n_k/x_k$, where n_k is the number of households selected in PSU k, and x_k is the total number of households in PSU k. That is, subsampling in the second stage does not depend on the first-stage samples and is carried out independently of subsampling in any other PSU. The probability of selecting PSU k in the second phase given a first-phase sample s_1 is $\pi_{k|s_1} = \sum_{s_2 \ni k} p_2(s_2|s_1) = \frac{\hat{x}_k}{a_{2,s_1}}$, where $\hat{x}_k$ is the estimated assumed total number of households in PSU k a_{2,s_1} is the sampling interval for selecting the second-phase-first-stage sample. This interval depends on the first-phase-first-stage sample. Every sample in the first phase has the same probability of being selected, i.e. $\pi_{s_1} = p_1(s_1) = 1/a_1$, where a_1 is the corresponding sampling interval. The inclusion probability of household h in PSU k can therefore be written as

$$\pi_{h,k} = \frac{n_k}{x_k}\frac{1}{a_1}\sum_{s_1 \ni k}\hat{x}_k\frac{1}{a_{2,s_1}}.$$

If, as was the case when subsample A was selected, the total number of households within PSU k has to be estimated, then a_1 and a_{2,s_1} depend on the estimates $\hat{x}_k$ which are generally not equal to their 'true' values. However, this is not the only problem of determining $\pi_{h,k}$. In fact, to determine $\pi_{h,k}$, for every s_1, the sampling intervals a_{2,s_1} had to be known, which is not the case.

Instead of determining weights based on the exact inclusion probabilities to calculate π-estimators (Horvitz and Thompson, 1952), one could determine weights to calculate the so-called π^*-estimators (see Särndal, Swensson, and Wretman, 1993, p. 347), which, as the π-estimators, can be shown to be unbiased for the corresponding population quantities. The probability $\pi^*_{h,k}$, which in general is different from $\pi_{h,k}$, can be written as

$$\pi^*_{h,k} = \pi_k \pi_{k|s_1} \pi_{h|s_1,s_2},$$

where π_k is the probability of selecting PSU k in the first phase of stage one and $\pi_{k|s_1}$ and $\pi_{h|s_1,s_2}$ are as defined above. Now, assuming x_k is known, $\pi_k = x_k/a_1$, $\pi_{k|s_1} = x_k/a_{2,s_1}$, $\pi_{h,k|s_1,s_2} = n_k/x_k$ and

$$\pi^*_{h,k} = \frac{x_k}{a_1}\frac{x_k}{a_{2,s_1}}\frac{n_k}{x_k}.$$

Unfortunately, still not all terms are known. A main problem is the determination of a_{2,s_1}. In phase two of the first stage, a program creating cells and filling them up with PSU's so as to mimic certain marginal distributions was used (see Infratest Sozialforschung, 1985, p. 106). Within each cell, systematic sampling was used. However, no information is available about how the program created these cells and which sampling intervals were used in the different cells. As for the determination of the exact inclusion probabilities, the number of households in each PSU had to be estimated. Given these problems, $\pi^*_{h,k}$ cannot be calculated.

To derive approximate weights, the following simplifying assumptions are made. Firstly, it is assumed that $\hat{x}_k \approx x_k$, so that

$$\tilde{\pi}^*_{h,k} = \frac{n_k \hat{x}_k}{a_1 a_{2,s_1}},$$

where $\tilde{\pi}^*_{h,k}$ denotes an approximation to $\pi^*_{h,k}$. It is further assumed that

$$\pi_{s_1} = 1/a_1 \approx \frac{\text{Number of PSU's in the first-phase sample}}{\text{Population total of households}},$$

which seems to be justified by the large number of private households in the target population ($N \approx 25007632$). Note that the integer part of the population total divided by the sampling interval is set to be equal to the number of PSU's in the first-phase sample. The same assumptions are made concerning the second-phase sample and furthermore, it is assumed that $1/a_{2,s_1}$ is approximately equal in every cell. Replacing $\hat{x}_k$ by $\hat{\bar{x}}_k$ and assuming that $\hat{\bar{x}}_k$ times the number of PSU's in the first-phase sample is approximately equal to the number of households in the first-phase sample and setting n_k equal to its estimated expected value (where n_k is considerd to be a random variable), then for every k, $\tilde{\pi}^*_{h,k}$ can be written as

$$\tilde{\pi}^*_h = \frac{\text{Number of private households selected}}{\text{Population total of private households}}.$$

However, although according to the target population non-private households had to be excluded, a few of them were selected. On the other hand, since the number of non-private households can only be identified for the observed portion of the sample, it is not possible to determine for each and every household in the selected sample whether it is a private or a non-private household. Therefore, only the identifiable non-private households are excluded. The number of private

households[1] in subsample A selected to be interviewed in 1984 was $n_{\mathrm{A}} = 7478$. The approximate weight is therefore given by $\tilde{\pi}_h^* = 7478/25007632$. Nothing can be derived for the non-private households. Therefore, in the absence of any information, they are given the same weight as the private households.

Note that $\hat{\pi}_h^*$ is approximately equal to the inclusion probability of household h if the second-phase selection were completely ignored (and the remaining assumptions were correct). Thus, as a model for determining the weight $w_h = 1/\hat{\pi}_h^*$, a two stage selection scheme is assumed, where the second phase of the first stage is ignored. Clearly, determination of the weight w_h rests on approximations derived from more or less plausible assumptions. It must again be noted that unfortunately, no information is available to derive weights avoiding some (or all) of the above assumptions.

Since the main arguments given above are similar for most of the other subsamples, only a rough description of the derivation of the weights for the other subsamples will be given.

4 Subsample B

The population from which subsample B was selected in 1983/1984 was defined to be the set of private households where the household head had the Turkish, Italian, Greek, (former) Yugoslavian or Spanish nationality. In fact, subsample B consists of five samples selected from the above five disjunct subpopulations. Each of the five subsamples was selected in two stages, where the first-stage samples were selected according to a systematic πps–scheme. The PSUs selected at the first stage were counties and metropolitan areas. The sizes of the PSUs were number of residents with the corresponding nationality. Given the first-stage samples of PSUs, within each PSU, addresses of persons aged 16 and older with a given nationality were selected according to a systematic sampling scheme with random start. The household selected in this manner was defined to be a sample element if the nationality of the household head was the same as the nationality

[1] The set of selected households also includes households which are a posteriori found to be non-private households. These households are identified only if they respond. Since only the observed non-private households are excluded, there may be a small number of non-private households in the unobserved part of the sample.

of the selected person.

The approximate inclusion probability of household h in PSU k for each of the five subsamples given this two stage design is

$$\begin{aligned}\pi_{h,k} &\approx (\text{Total number of persons living in household } h) \\ &\quad \times \frac{(\text{Number of persons selected in PSU } k)}{(\text{Total number of persons in PSU } k)} \\ &\quad \times \frac{(\text{Number of PSU's}) \times (\text{Total number of persons in PSU } k)}{(\text{Population total of persons})},\end{aligned}$$

where 'persons' means persons aged 16 and older and 'households' means private, valid households. Again, it is assumed that

$$\frac{1}{a_1} \approx \frac{(\text{Number of PSU's})}{(\text{Population total of persons})}$$

and

$$\frac{1}{a_2} \approx \frac{(\text{Number of persons selected in PSU } k)}{(\text{Total number of persons in PSU } k)}$$

and that estimates of quantities are approximately equal to their 'true' values. Similar to subsample A, replacing the number of selected persons in PSU k by its estimated expected value, we have for every k

$$\begin{aligned}\tilde{\pi}_h &= (\text{Total of persons living in household } h) \times (\text{Number of PSU's}) \\ &\quad \times \frac{(\text{Est. expected number of persons selected in PSU } k)}{(\text{Population total of persons})}.\end{aligned}$$

For the subsample of households where the head of the household had the Turkish nationality, $\tilde{\pi}_h$ is

$$\tilde{\pi}_h = \frac{(\text{Total of persons living in household } h) \times 80 \times 7.11}{965401}.$$

For the subsample of households where the head of the household had the (former) Yugoslavian nationality, $\tilde{\pi}_h$ is

$$\tilde{\pi}_h = \frac{(\text{Total of persons living in household } h) \times 40 \times 10.525}{444421}.$$

For the subsample of households where the head of the household had the Greek nationality, $\tilde{\pi}_h$ is

$$\tilde{\pi}_h = \frac{(\text{Total of persons living in household } h) \times 40 \times 7.475}{217304}.$$

For the subsample of households where the head of the household had the Italian nationality, $\tilde{\pi}_h$ is

$$\tilde{\pi}_h = \frac{(\text{Total of persons living in household } h) \times 40 \times 12.025}{441006}.$$

For the subsample of households where the head of the household had the Spanish nationality, $\tilde{\pi}_h$ is

$$\tilde{\pi}_h = \frac{(\text{Total of persons living in household } h) \times 40 \times 7.2}{137432}.$$

Again, nothing can be derived for the non-private households. Therefore, in the absence of any information, they are given the same weight as the private households.

5 Subsample C

Subsample C, selected in 1990, was a sample of private households in the former East Germany. The selection followed a 'two stage and two phases within the first stage' design, similar to the selection scheme used for subsample A. In the first phase of the first stage, communities (PSUs) were selected according to a systematic πps scheme with the sizes of the PSUs being the number of residents. The PSUs were then again stratified according to the variables used to sort the population elements so as to mimic certain marginal distributions. Within each cell, again samples of PSUs according to a systematic πps–scheme were selected. Given this second-phase-first-stage sample of PSUs, within each PSU, the households were selected according to a scheme that may approximately be described by a circular systematic sampling scheme with random start. The number of selected private households[2] in subsample C in 1990 was $n_{\mathrm{C}} = 3093$.

Since the scheme used to select subsample C is similar to the one used to select subsample A, the main arguments in deriving an approximate inclusion probability are similar as well[3]. Therefore, only the resulting approximate inclusion

[2] As in subsample A, there may be a few non-private households in this set of households.

[3] Note, however, that there are some slight differences. For example, the first-phase weights of the PSU's are the number of residents and not the number of households as in A. However, the assumptions necessary to derive the approximate inclusion probability are very similar to those needed to derive the weights for sample A households.

probability of household h is given, which is

$$\begin{aligned}\tilde{\pi}_h^* &= \frac{\text{Number of private households selected}}{\text{Population total of private households}} \\ &= \frac{3093}{5876672}\end{aligned}$$

(cf. Infratest Sozialforschung, 1992, p. 25).

Nothing can be derived for the non-private households. Therefore, in the absence of any information, they are given the same weight as the private households.

6 Subsample D

The target population can be defined as the set of private households with occupants who came to the former West Germany since 1984 but were not elements of the populations from which the samples A, B and C were selected. In fact, the part of D considered in this paper consists of two samples, selected in 1992/1994 (D_1) and 1994/1995 (D_2), respectively. As a result of several difficulties in selecting such a sample (for details, the reader is referred to Rendtel, Pannenberg and Daschke, 1997, or, Schulz et al. 1993), different selection shemes were used to select subsample D. In fact, one portion of sample D, D_1 selected in 1992 and 1994, consists of two subsamples, D_{11} and D_{12}, say, each selected according to a different selection scheme. The selection scheme of the other part of D, D_2 selected in 1994 and 1995, again differs from the selection schemes used to select the two subsamples D_{11} and D_{12}. However, the selection schemes of D_{11} and D_2 are similar in that the selection of the first-stage sample is based on a systematic πps–scheme. For both subsamples, the second-stage sample can approximately be described by a systematic sampling scheme with random start, where the valid sample elements are selected with a certain but unknown probability. Although this selection scheme has elements equal to the selection scheme used, e.g. for the selection of subsample A, there are also some differences. For example, selected households in 1992, as part of the D_{11} second-stage sample, were asked whether they agreed with the storage of their addresses for future surveys. These addresses then were used for selecting sample D_{11}. Given addresses selected in the same way in 1994, quota sampling elements were used to select sample D_2. The other part of D_1, D_{12}, was selected using telephone survey techniques, where phone

numbers were randomly choosen in view of regional criteria ('InfraScope' system, see Infratest Sozialforschung, 1994, or Rendtel, Pannenberg and Daschke, 1997). As for samples D_{11} and D_2, the selected households were then asked whether or not they agreed with the storage of their addresses for future surveys. Those who agreed were then selected in 1994 for subsample D_{12}.

From the design used, it is not possible to exactly determine the sample of selected private households used as a starting point to derive approximate weights. Calculation of the approximate inclusion probabilities in Rendtel, Pannenberg and Daschke (1997) is based on the set of observed households. Since in this paper the starting point for the determination of approximate weights are the selected private households and not the observed private households, the approximate inclusion probabilities are calculated in almost the same way as given in Rendtel, Pannenberg and Daschke (1997). That is, the derivation differs in that it does not use the estimate of the conditional *response* probability given that a household was selected and agreed with the storage of their addresses and, for sample D_2, given the quota sampling elements. Instead, it uses estimates of the conditional probabilities of being a *valid* household given it was selected and agreed with the storage of their addresses and, for sample D_2, given the quota sampling elements.

According to Rendtel, Pannenberg and Daschke (1997), the probability of being selected can be approximated by the sum of the probabilities of being selected for samples D_{11}, D_{12} or D_2, where the probabilities of being selected in two or all three samples are ignored. Then, the approximate inclusion probabilities are calculated as

$$\tilde{\pi}^*_h = \tilde{\pi}^*_{h,D_{11}} + \tilde{\pi}^*_{h,D_{12}} + \tilde{\pi}^*_{h,D_2}.$$

Three populations are distinguished, denoted as 'Übersiedler', 'Aussiedler' and 'Sonstige' (for details see Rendtel, Pannenberg and Daschke, 1997). The probability of being selected in sample D_{11} is equal for 'Übersiedler' and 'Aussiedler' and is approximated by

$$\tilde{\pi}^*_{h,D_{11}} = \frac{1}{7194} \times 0.787 \times \frac{172}{195},$$

where 1/7194 and 0.787 are derived in Rendtel, Pannenberg and Daschke (1997) and 172/195 is the proportion of valid households to the number of households

who agreed with the storage of their addresses (see Infratest Sozialforschung, 1994, p. 9). For 'Sonstige', $\tilde{\pi}^*_{h,D_{11}} = 0$.

The probability of being selected in sample D_{12} is equal for 'Übersiedler' and 'Aussiedler' and is approximated by

$$\tilde{\pi}^*_{h,D_{12}} = \frac{1}{7194} \times 0.444 \times \frac{172}{195},$$

where 1/7194 and 0.444 are derived in Rendtel, Pannenberg and Daschke (1997) and 172/195 is the same as above. Again, for 'Sonstige', $\tilde{\pi}^*_{h,D_{12}} = 0$.

The probability of being selected in sample D_2 is approximated by

$$\tilde{\pi}^*_{h,D_2} = \frac{1}{2687} \times 0.756 \times 0.527 \times \frac{385}{400}$$

for 'Übersiedler', where 1/2687, 0.756 and 0.527 are derived in Rendtel, Pannenberg and Daschke (1997) and 385/400 is the proportion of selected valid private households to selected households (see Infratest Burke Sozialforschung, 1996, p. 17). The probability for 'Aussiedler' is approximated by

$$\tilde{\pi}^*_{h,D_2} = \frac{1}{2687} \times 0.756 \times 0.696 \times \frac{385}{400},$$

where 1/2687, 0.756 and 0.696 are derived in Rendtel, Pannenberg and Daschke (1997) and 385/400 as above. The probability for 'Sonstige' is approximated by

$$\tilde{\pi}^*_{h,D_2} = \frac{1}{2687} \times 0.756 \times \frac{385}{400},$$

where 1/2687 and 0.756 are derived in Rendtel, Pannenberg and Daschke (1997) and 385/400 as above.

Therefore, the approximate household weights are given by the inverse approximate probabilities

$$\tilde{\pi}^*_h = \frac{(0.787 + 0.444) \times 172}{7194 \times 195} + \frac{0.756 \times 0.527 \times 385}{2687 \times 400} \approx 0.0002936$$

for 'Übersiedler',

$$\tilde{\pi}^*_h = \frac{(0.787 + 0.444) \times 172}{7194 \times 195} + \frac{0.756 \times 0.696 \times 385}{2687 \times 400} \approx 0.0003394$$

for 'Aussiedler' and

$$\tilde{\pi}^*_h = \frac{0.756 \times 385}{2687 \times 400} \approx 0.0002708$$

for 'Sonstige'. The approximate weights for a given subpopulation are identical regardless whether sample D is considered separately or in combination with all other subsamples.

7 Subsample E

In 1998, a new sample was selected from the population of households given by the union of the subpopulations as defined above. Note that this is in fact not the same population, but merely the same definition, since the population, of course, changes over time. The new sample, also denoted as subsample E, was selected independently from the ongoing panel (subsamples A through D). The selection scheme used for sample E essentially resembles the scheme also used to select subsample A. Again, the data are collected in two stages and two phases within the first stage, where the first- and second-phase samples are selected using the scheme also used for selecting subsample A. Although there are slight differences in the selection of the second-stage sample, mainly due to testing a new survey instrument (using a laptop for the personal interviews vs. paper-and-pencil personal interviews), the selection scheme is very similar to the one used to select the second-stage sample of subsample A. The number of selected private households[4] in subsample E in 1998 was $n_{\mathrm{E}} = 1969$ (see Infratest Burke Sozialforschung, 1998).

Since the scheme to select subsample E is very similar to the one used for subsample A, the derivation of the approximate weights is very similar too. Therefore, only the resulting weight ist given, which is

$$\begin{aligned} \tilde{\pi}_h^* &= \frac{\text{Number of private households selected}}{\text{Population total of private households}} \\ &= \frac{1969}{37571000}. \end{aligned}$$

As in subsamples A–D, non-private households are given the same approximate weights as the private households.

8 Subsample F

Subsample F was selected from the population of private households in Germany in 2000. Like subsample E, it was selected independently from all other subsamples and the selection schema was essentially the same as for selecting subsample A and F, however, with one exception. Like A and F, subsample E was selected in two stages and two phases within the first stage. The difference between F

[4] As in subsample A, there may be a few non-private households in this set of households.

and subsamples A and E was in the selection of households within PSU's as follows. First, the population was divided in to two parts: Those households with all adults (age ≥ 16) owing the german nationality ('german' households) and those households where at least one adult has not the german nationality ('non-german' households). Within each PSU, 24 households were selected according to the same schema as in subsamples A and E. However, 'german' households were selected mainly using the first 12 adresses within each PSU. In fact, a few were selected from the second 12 adresses as well. The 'non-german' households were selected using all the 24 adresses.

As for subsample A, the following assumptions are made. First, it is assumed that $\hat{x}_k \approx x_k$,

$$\pi_{s_1} = 1/a_1 \approx \frac{\text{Number of PSU's in the first-phase sample}}{\text{Population total of households}}$$

and that $\bar{\hat{x}}_k$ times the number of PSU's in the first phase sample is approximately equal to the number of households in the first-phase sample, where $\hat{x}_k \approx \bar{\hat{x}}_k \ \forall k$. Then the approximate probabilities $\pi^*_{h,k}$ can be written as

$$\pi^*_{h,k} = n_k \frac{\text{Number of PSU's in the second-phase sample}}{\text{Population total of households}}.$$

An approximation to n_k was calculated as follows. First, let $n_{k,g}$ and $n_{k,\bar{g}}$ be the number of selected valid adresses of 'german' households and 'non-german' households, respectively. Then for the first 12 adresses, n_k was identical for both, the 'german' and the 'non-german' households, i.e. 24 minus those adresses that turned out to be invalid, e.g. unoccupied apartments or houses. However, even if the households could be interviewed, some turned out to be non-private households, e.g. dormitories. Therefore, n_k had to be reduced by an estimate of the number of non-private households from the sample of valid households. Another problem is the number of selected households from the second part of the sample of the 24 adresses. For this part, n_k had to be estimated from the first part of the sample. For both, the 'german' households and the 'non-german' households this was done by multiplying the mean number of private 'german' or 'non-german' households with the ratio of the mean of n_k from the the 'first-part' sample and the mean of the number of 'german' and 'non-german' households, respectively, where means were taken over the second-phase PSU's. Then, the approximation of $n_{k,g}$ is given by

$$
\begin{aligned}
n^*_{k,g} = & \frac{\text{No. of observed private households in } s_{1,k}}{\text{No. of observed households in } s_{1,k}} \\
& \times \text{No. of selected valid households in } s_{1,k} \\
& + \text{Mean over } k \text{ of observed private 'german' households in } s_{2,k} \\
& \times \frac{\text{Mean over } k \text{ of est. selected private valid households in } s_{1,k}}{\text{Mean over } k \text{ of observed private 'german' households in } s_{1,k}},
\end{aligned}
$$

where $s_{1,k}$ and $s_{2,k}$ denote the 'first-part' and the 'second-part' sample in PSU k, respectively. Correspondingly, for the 'non-german' households we have

$$
\begin{aligned}
n^*_{k,\bar{g}} = & \frac{\text{No. of observed private households in } s_{1,k}}{\text{No. of observed households in } s_{1,k}} \\
& \times \text{No. of selected valid households in } s_{1,k} \\
& + \text{Mean over } k \text{ of observed private 'non-german' households in } s_{2,k} \\
& \times \frac{\text{Mean over } k \text{ of est. selected private valid households in } s_{1,k}}{\text{Mean over } k \text{ of observed private 'non-german' households in } s_{1,k}}.
\end{aligned}
$$

Replacing $n^*_{k,g}$ and $n^*_{k,\bar{g}}$ by their respective means over the PSU's, $\bar{n}^*_{k,g} \approx 11.00$ and $\bar{n}^*_{k,\bar{g}} \approx 19.55$, and since the number of PSU's in the second phase sample is 985 and the number of private households in Germany in the year 2000 was 38124000, whe have

$$
\pi^*_{h,k} = \bar{n}_k \, \frac{985}{38124000},
$$

where $\bar{n}_k$ is $\bar{n}_{k,g}$ for 'german' household and $\bar{n}_{k,\bar{g}}$ for 'non-german' households. Nothing, can be said about non-private households. Therefore, they received the same weights as the private households.

9 Concluding Remarks

From a formal point of view, the weights derived in this paper cannot be interpreted as inclusion probabilities of the selected elements (households). On the other hand, given the information about the sampling schemes, it seems impossible to derive in some sense 'better' weights for every possible analysis to be made with the GSOEP. For special problems at hand, there may of course exist better solutions (e.g. using regression estimators). However, they clearly have to

be worked out by an analyst himself/herself for the specific estimation problem given.

It should be noted that the weights derived in this paper are not intended to replace the standard weights supplied with the GSOEP disk. Merely, the approximate design weights derived in this paper are thought as to supplement the standard weights in that they may form the basis for the calculation of alternative weights by modeling e.g. nonresponse or attrition mechanisms in a different way than it is done for the standard weights (e.g. Rendtel, 1995).

APPENDIX

Table 1 and 2 give the values of the approximate inclusion probabilities derived in sections 3 – 7. The inverse of these approximate inclusion probabilities are stored under the name DESIGN in file HHRF on the GSOEP CD.

Table 1: *Approximate inclusion probabilities for subsamples A, C, D, F and E* ($\times 10^{-4}$)

Subsample							
A	C	D_1	D_2	D_3	E	F_1	F_2
.2990	.5263	.2936	.3394	.2708	.0524	.284	.505

D_1: 'Übersiedler' D_2: 'Aussiedler' D_3: 'Sonstige'. F_1: All adult (age ≥ 16) household members have the german nationality F_2: At least one of the adult (age ≥ 16) household members has not the german nationality.

Table 2: *Approximate inclusion probabilities for subsample B_1 – B_5* ($\times 10^{-4}$)

Sub-sample	Number of household members (≥ 16 years)						
	1	2	3	4	5	6	7
B_1	0.5892	1.1784	1.7676	2.3567	2.9459	3.5351	
B_2	0.9473	1.8946	2.8419	3.7892	4.7365		
B_3	1.3760	2.7519	4.1279	5.5038	6.8798		
B_4	1.0907	2.1814	3.2721	4.3628	5.4534		
B_5	2.0956	4.1912	6.2867	8.3823	10.478		14.669

B_1: Turkish, B_2: (former) Yugoslavian, B_3: Greek, B_4: Italian, B_5: Spanish nationality of the household head.

References

Horvitz. D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47,* 663–685.

Infratest Sozialforschung (1985). *Das Sozio-ökonomische Panel, Welle 1, Methodenbericht zur Haupterhebung.* München: Infratest Sozialforschung.

Infratest Sozialforschung (1992). *Das Sozio-ökonomische Panel in Ostdeutschland, 1990-1991, Methodenbericht zur Haupterhebung, 1-2 Ost.* München: Infratest Sozialforschung.

Infratest Sozialforschung (1994). *SOEP '94, Zuwanderer-Befragung (Teilstichprobe D 1), Methodenbericht.* München: Infratest Sozialforschung.

Infratest Burke Sozialforschung (1996). *SOEP '95, Zuwandererbefragung II, Zweitbefragung D 1, Erstbefragung D 2, Methodenbericht.* München: Infratest Burke Sozialforschung.

Infratest Burke Sozialforschung (1998). *SOEP '98, Erstbefragung der Stichprobe E, Methodenbericht.* München: Infratest Burke Sozialforschung.

Pannenberg, M., Pischner, R., Rendtel, U. & Wagner, G.G. (1998). Sampling and Weighting. In: J.P. Haisken-De New & J.R. Frick, *Desktop Companion to the German Socio-Economic Panel Study (GSOEP), Version 2.2 (chap. 4).* Berlin: German Institute for Economic Research.

Rendtel, U. (1995). *Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität.* Frankfurt: Campus.

Rendtel, U., Pannenberg, M. & Daschke, S. (1997). Die Gewichtung der Zuwanderer-Stichprobe des Sozio-oekonomischen Panels (SOEP). *Vierteljahreshefte zur Wirtschaftsforschung, 66, 2,* 271–286.

Särndal, C.-E., Swensson, B. & Wretman, J. (1993). *Model assisted survey sampling.* New York: Springer.

Schulz, E., Rendtel, U., Schupp, J. & Wagner, G. (1993). *Das Zuwanderer-Problem in Wiederholungsbefragungen am Beispiel des Sozio-oekonomischen Panels (SOEP).* Deutsches Institut für Wirtschaftsforschung, Diskussionspapier Nr. 71.

Wagner, G., Schupp, J. & Rendtel, U. (1994). Das Sozio-ökonomische Panel – Methoden der Datenproduktion und -aufbereitung im Lngsschnitt. In R. Hauser, N. Ott & G. Wagner (Hrsg.), *Mikroanalytische Grundlagen der Gesellschaftspolitik – Band 2, Erhebungsverfahren, Analysemethoden und Mikrosimulation* (pp. 70–112). Berlin: Akademie Verlag.