Rendtel, Ulrich

The effect of panel attrition on the variance of population estimates from household panels

DIW Discussion Papers, No. 81

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Rendtel, Ulrich (1993) : The effect of panel attrition on the variance of population estimates from household panels, DIW Discussion Papers, No. 81, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/95716

DIW Diskussionspapiere
Discussion Papers

Discussion Paper No. 81

The Effect of Panel Attrition on the Variance of Population Estimates from Household Panels

by
Ulrich Rendtel*

Deutsches Institut für Wirtschaftsforschung, Berlin
German Institute for Economic Research, Berlin

Deutsches Institut für Wirtschaftsforschung

Discussion Paper No. 81

The Effect of Panel Attrition on the Variance of Population Estimates from Household Panels

by
Ulrich Rendtel*

*) German Institute for Economic Research (DIW), Berlin

Berlin, November 1993

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 14191 Berlin
Telefon: 49-30 - 82 991-0
Telefax: 49-30 - 82 991-200

The Effect of Panel Attrition on the Variance of Population Estimates from Household Panels*

Ulrich Rendtel
German Institute for Economic Research
Berlin

November 1993

Abstract

Panel attrition has not only the potential to bias population estimates but it may also inflate the variance of the estimates from panel surveys. Thus it is essential for an ongoing panel survey to monitor not only the size of the panel attrition and the potential biases that may occur but also the decrease in the precision of estimates. Losses in the precision of cross-sectional estimates are also due to a second source, which is specific to the design of household panels. This effect stems from different inclusion probabilities of households depending on the number of occasions when persons move into the household.

Panel surveys offer substantial efficiency gains in trend analysis, where differences in time with respect to population charcteristics are estimated. Such an analysis may also be performed on the basis of a sequence of independent cross-sectional surveys. It is investigated here, how the panel attrition effects the efficiency gains of the panel compared to a sequence of independent cross–sections.

*Paper presented at the Final Conference of the European Science Foundation Network on Household Panels in Luxemburg

Panel attrition also offers the possibility of different population estimates for the same population characteristic. It is investigated which of the alternative estimates is more precise.

In order to give an answer to the above topics two panel specific approaches of variance estimation are presented. Numerical examples are displayed from the German Socio–economic Panel (SOEP).

Key words:Sampling variance, household panel, panel attrition.

1 Introduction

A household panel survey differs from a series of independent cross-sectional surveys in that each wave of the panel provides the basis for the following wave of interviewing. In a household panel some losses occur in the following wave: people die or leave the sampling area (demographic losses) or they declare themselves unwilling to participate in the next wave (panel attrition). Also gains occur in the sample : new persons enter the panel households or panel persons leave their households and move together with other persons, which are also interviewed. Also children in panel households reach the age at which they are eligible to be interviewed.

It is immediately obvious that such special characteristics of the survey have to be taken account of in the weighting procedure. In a companion paper Rendtel (1993b) has shown how the method of inverse probalitity weighting can be applied to a household-panel and what difficulties have to be overcome in doing this work. If the models for the non-response and the inclusion of new persons are correctly specified one gets different different weighting schemes for cross-sections and longitudinal tabulations that produce unbiased population estimates.In the case of a panel there are also some possibilities to check the model specifications for non-response, cf. Frick/Rendtel/Wagner (1993).

The present paper deals with the effects of panel attrition and design effects of a household panel on the variance of population estimates.

There are at least three panel specific reasons to estimate the variance of population estimates:

- Panel attrition may not only bias the population estimates but it may also — by differential non-response — inflate the variance of population estimates.

- A panel promises to give more accurate estimates of differences in time with respect to population totals than a series of independent cross-sections. This is seen from:

$$V(\hat{T}_{Y_2} - \hat{T}_{Y_1}) = V(\hat{T}_{Y_2}) + V(\hat{T}_{Y_1}) - 2C(\hat{T}_{Y_1}, \hat{T}_{Y_2})$$

  where $\hat{T}_{Y_t}$ estimates the population total $T_{Y_t} = \sum_{i \in G_t} Y_{i,t}$ at time t. In a panel $C(\hat{T}_{Y_1}, \hat{T}_{Y_2})$ is usually positive and causes a large efficieny gain

over cross-sectioal surveys and other sampling designs with overlapping samples,cf. for example Särndal et al. (1992, p. 377). The size of gain depends on the stability of $Y_{i,t}$ with respect to time. Rendtel(1991) calculates the gain that is achieved by $C(\hat{T}_{Y_1}, \hat{T}_{Y_2})$ under the restriction that the panel attrition may be ignored. Thus it important to see how these benefits from the panel design have to be balanced against the effects from panel attrition.

- Panel attrition creates differences between estimators which are algebraically equivalent under no attrition. Suppose one wants to estimate $D = T_{Y_2} - T_{Y_1}$, the difference of the population totals at time 1 and 2. If there is no panel attrition D is efficiently estimated by the difference of the cross-sectional estimates $\hat{D}_C = \hat{T}_{Y_2} - \hat{T}_{Y_1}$. Because of

$$\begin{aligned} D &= \sum_{G_2} Y_{i,2} - \sum_{G_1} Y_{i,1} \\ &= \sum_{G_2 \cap G_1} (Y_{i,2} - Y_{i,1}) + \sum_{G_2 \setminus G_1} Y_{i,2} - \sum_{G_1 \setminus G_2} Y_{i,1} \\ &= \sum_{G_2 \cap G_1} d_i + G_+ - G_- \\ &= G_{1,2} + G_+ - G_- \end{aligned}$$

one may estimate each of these components separately. Hence $d_i = Y_{i,2} - Y_{i,1}$ is different from 0 only in the case where the characteristic of interest changes. The component $G_{1,2}$ may be estimated from all persons that are in wave 1 and in wave 2. The observations are weighted by the inverse longitudinal inclusion probabilities $\pi_{i,1,2}$. Thus one gets:

$$\hat{D}_L = \sum_{S_1 \bigcap S_2} \frac{1}{\pi_{i,1,2}} d_i + \sum_{S_2 \setminus G_1} \frac{1}{\pi_{i,2}} Y_{i,2} - \sum_{S_1 \setminus G_2} \frac{1}{\pi_{i,1}} Y_{i,1}$$

where the last two expressions estimate G_+ and G_-. If $\pi_{i,1} = \pi_{i,2} = \pi_{i,1,2}$, which implies that there is no panel attrition, it holds that:

$$\begin{aligned} \hat{D}_L &= \sum_{S_2} \frac{1}{\pi_{i,2}} Y_{i,2} - \sum_{S_1} \frac{1}{\pi_{i,1}} Y_{i,1} \\ &= \hat{D}_Q \end{aligned}$$

If $\hat{D}_C$ and $\hat{D}_L$ are different, the question arises, which estimator is more precise.

A special feature of a housedold panel is the cluster effect of sampling all persons within a sampled household. If the analysis is done on an individual level this introduces a substantial variance effect, which depends on the homogenity of the characteristic within households. On the other hand if the interest of the survey is on interactions of individual behaviour within households there is no way to avoid this clustering.

This clustering of the sample in the first panel wave is carried over to subsequent waves by the follow-up. Two opposing tendencies arise. On the one hand, the clusters become larger by people joining existing households during the panel. On the other hand, there are households which split off. In these cases the characteristics within the cluster become progressively inhomogeneous. While the enlargement of clusters has the potential to increase the variance, the decrease in homogeneity within the cluster implies a decrease in variance.

In order to give an answer to the above questions, one has to be able to compute reasonable estimates for the variance of these population estimates. Therefore the paper presents two approximations for the variance of cross–sectional and longitudinal population estimates, which reflect carefully the sampling design of a household panel. Thus the approach presented here is also design–oriented (cf. Rendtel 1993b) and no attempt is made here to derive model–based variance estimates; see Rao/Wu (1988) and Sitter (1992) for model–based replication methods of variance estimation.

Numerical examples are displayed from German **SO**cio-**E**conomic **P**anel (SOEP). A detailed description of this household panel, which started in 1984 on the basis of 6000 households, can be found in Hanefeld (1984) and Wagner et al. (1992).

2 Bounds to the variance of population estimates

In order to derive bounds to the variance of population estimates from a household panel we assume:

AS 1: All household members of a sampled household belong to the sample at the individual level.

AS 2: All interviewed persons will be followed into the next panel wave.

AS 3: The panel participation is uniform within households.

AS 4: The success of follow-up is positively correlated among persons; i. e.:

$$\pi_{P_i,P_j} \geq \pi_{P_i}\pi_{P_j}$$

where π_{P_i} (resp. π_{P_j}) is the probability of successful follow-up of person i (resp. person j).

AS 5: The selection of wave-1-households is independent.

Assumptions 1 to 3[1] guarantee that the inclusion probabilities of households and its household members are identical for cross-sections. Empirical evidence from the SOEP suggests that assumption 3 is almost perfectly met, cf. Rendtel (1990,1993a). Assumption 4 states that the willingness to participate in later panel waves is not negatively correlated. Note that in a household panel it will happen that persons living in different households know of each others participation in the survey. This happens, if two persons split off from the same sampled household during the panel. So there may be a potential interaction in their participation behaviour. The empirical results from the SOEP suggest that there are such interactions and that they are positively correlated; cf. Rendtel (1993a).

Assumption 5 guarantees (together with assumptions 2,3 and 4) positive joint inclusion probabilities for all persons and households in all panel waves. This is a necessary condition for the variance estimation by inverse joint inclusion probabilities. Assumption 5 has also important consequences for the set of units (persons or households) with dependent selection, which will be discussed later.

One has to admit that assumption 5 is a simplifiing condition which is not met by the SOEP sampling design. The first wave sampling of the SOEP households used two-stage sampling. On each stage systematic sampling was

[1]Note that assumptions 1 and 2 are not met by the Panel Study of Income Dynamics (PSID), cf. Hill (1992).

in action to select the primary sampling units (PSU's) and the secondary sampling units (SSU's), i.e. the households. A description of the SOEP sampling procedures is given in appendix A. The systematic sampling results in joint inclusion probabilities $\pi_{i,j}$, which are 0 for neighbouring units i and j, and causes effects negative covariance terms in

$$V(\hat{T}_Y) = \sum_{i \in G}(\frac{1}{\pi_i} - 1)Y_i^2 + \sum_{i \in G}\sum_{j \neq i}(\frac{\pi_{i,j}}{\pi_i \pi_j} - 1)Y_i Y_j \quad (1)$$

where $\hat{T}_Y = \sum_{i \in S} \frac{1}{\pi_i} Y_i$ is the inverse probability estimator of $T_Y = \sum_{i \in G} Y_i$. Thus in the case of the SOEP, variance estimates on the basis of AS 5 tend to overestimate $V(\hat{T}_Y)$. Therefore a different approach of variance estimation is presented in section 3, which reflects the use of the systematic sampling during the first panel wave.

The variance estimation in this section bases on:

$$\hat{V}(\hat{T}_Y) = \sum_{i \in S}(\frac{1}{\pi_i} - 1)\frac{Y_i^2}{\pi_i} + \sum_{i \in S}\sum_{j \in D_i \cap S}(\frac{\pi_{i,j}}{\pi_i \pi_j} - 1)\frac{Y_i Y_j}{\pi_{i,j}} \quad (2)$$

where D_i is the set of units (persons or households) with $\pi_{i,j} \neq \pi_i \pi_j$. Hence D_i is the set of units with dependent selection with respect to unit i. In following subsections we will determine the set D_i for cross-sectional and longitudinal samples. Lower und upper bounds for $\hat{V}(\hat{T}_Y)$ may be derived by lower und upper bounds for $(\frac{\pi_{i,j}}{\pi_i \pi_j} - 1)/\pi_{i,j}$.

2.1 Estimates of cross-sectional totals

Let $Y_{i,t}$ be the characteristic of interest in wave t, $\pi_{i,t}$ the cross-sectional inclusion probability of unit i and $\hat{T}_{Y_t} = \sum_{i \in S_t} \frac{1}{\pi_{i,t}} Y_{i,t}$ the population estimate of T_{Y_t}.

The variance of $\hat{T}_{Y_t}$ is given by:

$$V(\hat{T}_{Y_t}) = \sum_{i \in G_t}(\frac{1}{\pi_{i,t}} - 1)Y_{i,t}^2 + \sum_{i \in G_t}\sum_{j \in D_{i,t}}(\frac{\pi_{i,j,t}}{\pi_{i,t}\pi_{j,t}} - 1)Y_{i,t}Y_{j,t} \quad (3)$$

where G_t is the cross-sectional universe at time t and $D_{i,t}$ is the set of units whose choice does depend on the choice of unit i at wave t.

If the joint inclusion probabilities of all units in G_t are positive, $V(\hat{T}_Y)$ may be unbiased estimated by:

$$\hat{V}(\hat{T}_{Y_t}) = \sum_{i \in S_t} (\frac{1}{\pi_{i,t}} - 1) \frac{Y_{i,t}^2}{\pi_{i,t}} + \sum_{i \in S_t} \sum_{j \in D_{i,t} \cap S_t} (\frac{\pi_{i,j,t}}{\pi_{i,t}\pi_{j,t}} - 1) \frac{Y_{i,t} Y_{j,t}}{\pi_{i,j,t}} \quad (4)$$

Because of assumptions 1 to 5 the selection of persons is dependent if the selection of their households is dependent in wave t.

Denote by $F_{h,t}$ the set of all wave–1–households in G_t that contribute to the selection of household h in wave t by the follow–up rules. It is not important here wether the follow–up was successful or not. $F_{h,t}$ consists of those households, where household h can be reached in wave t by the follow–up rules of the panel.

The follow–up rules induce a tree structure among the households of the preceding waves. The root of the tree is household h at wave t. This root divides into branches at some wave $t'(1 \leq t' < t)$ if follow–up persons from different households move into the household. The branches are the households of the follow–up persons and the original household. This scheme is repeated for each branch until wave 1 is reached. For households with no move–in's the tree is deteriorated and has only the root.

The inclusion probability $\pi_{h,t}$ in wave t is given by the inclusion probability of the tree households $F_{h,t}$ in wave 1 multiplied by the probability of their successful follow–up, cf. Rendtel (1993b) for details. Thus the larger the number of move–in's into a panel household the larger is its inclusion probability $\pi_{h,t}$.

Under assumptions 1 to 5 the selection of households h and k in wave t is dependent if:

$$F_{h,t} \bigcap F_{k,t} \neq \emptyset \quad (5)$$

If $F_{h,t} \bigcap F_{k,t} = \emptyset$ there is no interaction with respect to follow–up, since persons in these households don't know of their participation in the panel. Hence the independence of selection follows from the independent choice of the wave–1–households.

Suppose for a moment that the panel attrition may be ignored. If $F_{h,t} \bigcap F_{k,t} \neq \emptyset$ then:

$$\pi_{h,t} \leq \pi_{h|k,t} \leq 1 \tag{6}$$

where $\pi_{h|k,t}$ is the selection probability of household h conditional that household k was selected. If the inclusion probabilities of wave–1–households are approximately equal, $\pi_{h|k,t}$ is given by the ratio:

$$\pi_{h|k,t} \approx \frac{\text{no. of households in } F_{h,t} \cap F_{k,t}}{\text{no. of households in } F_{k,t}} \tag{7}$$

Thus $\pi_{h|k,t}$ may be assumed to be much larger than $\pi_{h,t}$.

If panel attrition is to be accounted for, the design probabilities have to be multiplied by the probabilities of successful follow–up. While on the left side of eqn. (6) the unconditional probabilities of successful follow–up have to be used, on the right side of eqn. (6) the conditional probabilities of follow–up have to be used. Because of the assumed positive correlation of follow–up the conditional probabilities of follow–up are increased by the information that household k was successfully selected in wave t. Hence eqn. (6) holds also in the case of panel attrition.

Since the inclusion probabilities for households and household members are identical, eqn. (6) may be extended at the individual level:

$$\pi_{i,t} \leq \pi_{i|j,t} \leq 1 \tag{8}$$

Therefore all summands in the covariance term of $\hat{V}(\hat{T}_{Y_t})$ are non–negative. Hence a lower bound $\hat{V}_L$ for $\hat{V}(\hat{T}_{Y_t})$ is given by replacing $D_{i,t}$ by a smaller set. Denote by $H_{i,t}$ the set of persons that live together with person i in a household at wave t. For $j \in H_{i,t}$ it holds:

$$\pi_{i,j,t} = \pi_{i,t} = \pi_{j,t} \tag{9}$$

Hence

$$\hat{V}_L = \sum_{i \in S_t} (\frac{1}{\pi_{i,t}} - 1) \frac{Y_{i,t}^2}{\pi_{i,t}} + \sum_{i \in S_t} \sum_{j \in H_{i,t} \cap S_t} (\frac{1}{\pi_{i,t}} - 1) \frac{Y_{i,t} Y_{j,t}}{\pi_{j,t}} \tag{10}$$

is a lower bound for $\hat{V}(\hat{T}_{Y_t})$.

An upper bound for $\hat{V}(\hat{T}_{Y_t})$ is achieved when $\pi_{i|j,t} = \pi_{i,j,t}/\pi_{j,t}$ is replaced by its upper bound 1. This yields:

$$\hat{V}_U = \sum_{i \in S_t} (\frac{1}{\pi_{i,t}} - 1) \frac{Y_{i,t}^2}{\pi_{i,t}} + \sum_{i \in S_t} \sum_{j \in D_{i,t} \cap S_t} (\frac{1}{\pi_{i,t}} - 1) \frac{Y_{i,t} Y_{j,t}}{\pi_{j,t}} \tag{11}$$

Let us see wether there is a simple characterization or the set $D_{i,t} \cap S_t$. Again we use a tree structure on the set of households. But this tree structure is different from the previous structure which started in wave t and generated the set $F_{h,t}$. This tree structure uses the natural time ordering and starts in wave 1. The set $G_{h,1}$ includes only the household h, which is the root of the tree. $G_{h,2}$ consists of those wave–2–households that are generated from household h by splitting–off[2]. $G_{h,3}$ is the set of wave–3–households that is generated by the splitting of the $G_{h,2}$ households. In this manner $G_{h,t}$ consists of those households[3], which are generated by splitting–off from the wave–1–household h.

There is a positive possibility that two such trees grow together. This may happen if some panel persons from different trees move together. But the probability of such an event is rather small, if the selection of wave–1–households is independent and rather low[4]. During the first 8 waves of the SOEP no such case occurred. Thus for every household $h \in S_t$ there is a unique root household R(h,t) such that $h \in G_{R(h,t),t}$.

This definition may be extended to the individual level. Thus $R(i,t) = R(h(i),t)$ is the root household of the household h(i), where person i lives at wave t. This root household $R(i,t)$ generates the set $G_{R(i,t),t}$ of wave–t–households that can be reached from household $R(i,t)$ by follow–up. The set of persons in the $G_{R(i,t),t}$ households will be denoted by $G_{i,t}$.

[2]This includes also the "old" household h

[3]Including household h

[4]The inclusion probability of SOEP households is about 1/4500 on the average.

We will now show that $D_{h,t} \cap S_t$ is given by $G_{R(h),t} \cap S_t$. For $k \in G_{R(h),t}$ there is some household l that originated from household $R(h)$. By tracing the roots of houshold l one will find after some steps that $R(h)$ belongs to $F_{k,t}$. By definition $R(h)$ belongs also to $F_{h,t}$. Hence $F_{h,t} \cap F_{k,t} \neq \emptyset$ and $k \in D_{h,t}$. To show the opposite direction we assume $k \in D_{h,t} \cap S_t$. Hence $\Delta = F_{h,t} \cap F_{k,t}$ is not empty. Suppose that $\Delta \cap S_1$ has more than one element. Then it is possible to reach the households h and k from two different wave–1–households of S_1. This is in contradiction with the assumption that there exists only one unique root household. If $\Delta \cap S_1$ is empty there is no possibility to reach both households by follow–up of the wave–1–households in S_1. Thus $\Delta \cap S_1$ has a unique element l. Hence $R(h) = R(k) = l$ and $k \in G_{R(h),t} \cap S_t$.

The result $D_{h,t} \cap S_t = G_{R(h),t} \cap S_t$ is immediately extended to the individual level. This yields:

$$D_{i,t} \cap S_t = G_{i,t} \cap S_t$$

The above representation of households and persons with dependent selection guarantees a simple computation of $\hat{V}_U$ [5] and has also an appealing interpretation. While the lower bound $\hat{V}_L$ refers to the cluster units at the start of the panel, $\hat{V}_U$ refers to the "clusters" that are generated by the follow–up of the original households.

In households with no split–off's the covariance contribution in $\hat{V}_L$ and $\hat{V}_U$ will coincide. Hence the difference between $\hat{V}_L$ and $\hat{V}_U$ will increase with the number of households with split–off's. This percentage will increase with the number of panel waves. Yet in wave 8 of the SOEP the percentage of households with split–off's is still quiete low[6]. So one will expect only small differences between $\hat{V}_L$ and $\hat{V}_U$.

2.2 Estimates of longitudinal totals

Longitudinal tabulations from wave 1 to wave t refer to a longitudinal universe $G = \cap_{\tau=1,\ldots,t} G_\tau$. The tabulations are taken from persons that participate from wave 1 until wave t. Thus the longitudinal sample S is given

[5] The sets $G_{R(h),t} \cap S_t$ and $G_{i,t} \cap S_t$ are characterized in the SOEP data base by the upmost data–base key.

[6] Until wave 8 only 18.2% of wave–1 households splitted into different households. This percentage refers to those cases where the wave–1 household gave an interview in wave 8 and also at least one split–off household is in the sample at wave 8

by the intersection of the cross-sectional samples, i.e. $S = \bigcap_{\tau=1,\ldots,t} S_\tau$ The longitudinal inclusion probability $\pi_{i,1,t}$ is given by the product of the cross-sectional inclusion probability $\pi_{i,1}$ and the probability of successful follow-up from wave 1 to wave t; see Rendtel (1993b) for details.

The joint longitudinal inclusion probabilities $\pi_{i,j,1,t}$ are given by the product of $\pi_{i,j,1}$ and the probability of joint follow-up from wave 1 to wave t.

We first deal with the case that persons i and j live together in wave 1. Let $t_{i,j}$ be the last panel wave where persons i and j live together. If they don't separate until wave t, we define $t_{i,j} = t$. In this case we get:

$$\begin{aligned} \pi_{i,j,1,t} &= \pi_{i,1,t_{i,j}} \pi_{P_i,P_j} \\ &= \pi_{i,1,t} \pi_{P_j|P_i} \end{aligned} \tag{12}$$

where π_{P_i,P_j} is the joint probability of successful follow-up after wave $t_{i,j}$ and $\pi_{P_j|P_i}$ is the conditional probability of successful follow-up of person j if person i is followed successfully. Assumption 4 implies:

$$\pi_{P_j} \leq \pi_{P_j|P_i} \leq 1 \tag{13}$$

Now we treat the case that person i and person j live in different households in wave 1. Hence they live in households that do not come into contact as long as their trees $G_{i,t}$ and $G_{j,t}$ keep separate. In this case their choice is independent (due to AS 5) and also is their participation behaviour. It was pointed out that in a household panel probably no other cases occur. Hence the set $D_{i,t} \cap S$ of persons which contribute to the covariance term reduces to those persons in S, that lived at wave 1 together with person i.

If Y_i is the longitudinal characteristic of interest and if T_Y denotes the total in the longitudinal universe, $\hat{T}_Y = \sum_{i \in S} \frac{1}{\pi_{i,1,t}} Y_i$ is the inverse probability estimator of T_Y. Its variance is given by:

$$V(\hat{T}_Y) = \sum_{i \in G} \left(\frac{1}{\pi_{i,1,t}} - 1\right) Y_i^2 + \sum_{i \in G} \sum_{j \in D_{i,t}} \left(\frac{\pi_{i,j,1,t}}{\pi_{i,1,t}\pi_{j,1,t}} - 1\right) Y_i Y_j \tag{14}$$

This variance may be estimated by :

$$\hat{V}(\hat{T}_Y) = \sum_{i\in S}(\frac{1}{\pi_{i,1,t}}-1)\frac{Y_i^2}{\pi_{i,1,t}} + \sum_{i\in S}\sum_{j\in H_{i,1}\cap S}(\frac{\pi_{i,j,1,t}}{\pi_{i,1,t}\pi_{j,1,t}}-1)\frac{Y_iY_j}{\pi_{i,j,1,t}}$$

$$= \sum_{i\in S}(\frac{1}{\pi_{i,1,t}}-1)\frac{Y_i^2}{\pi_{i,1,t}} + \sum_{i\in S}\sum_{j\in H_{i,1}\cap S}(\frac{1}{\pi_{j,1,t}}-\frac{1}{\pi_{P_j|P_i}})\frac{Y_iY_j}{\pi_{i,1,t}} \quad (15)$$

If we use the lower and upper bounds for $\pi_{P_j|P_i}$ in eqn. (13) we get lower and upper bounds for $\hat{V}(\hat{T}_Y)$:

$$\hat{V}_L = \sum_{i\in S}(\frac{1}{\pi_{i,1,t}}-1)\frac{Y_i^2}{\pi_{i,1,t}} + \sum_{i\in S}\sum_{j\in H_{i,1}\cap S}(\frac{1}{\pi_{j,1,t}}-\frac{1}{\pi_{P_j}})\frac{Y_iY_j}{\pi_{i,1,t}} \quad (16)$$

$$\leq \hat{V}(\hat{T}_Y)$$

$$\leq \sum_{i\in S}(\frac{1}{\pi_{i,1,t}}-1)\frac{Y_i^2}{\pi_{i,1,t}} + \sum_{i\in S}\sum_{j\in H_{i,1}\cap S}(\frac{1}{\pi_{j,1,t}}-1)\frac{Y_iY_j}{\pi_{i,1,t}}$$

$$= \hat{V}_U \quad (17)$$

The lower bound is achieved if the participation behaviour of persons i and j is independent after they separated to different households.

In cases where person i and j live together until wave t we have $\pi_{P_j} = 1$. Thus in these cases the covariance contribution in $\hat{V}_L$ and $\hat{V}_U$ is identical. Hence differences between $\hat{V}_L$ and $\hat{V}_U$ depend on the number of persons that separate from their wave–1–households. This percentage will increase during the panel. In the SOEP however the percentage is still low [7].

The computation of $\hat{V}_L$ makes it necessary to use π_{P_j}, the probability of successful follow-up of person j after wave $t_{i,j}$. This is computationally unattractive since for each pair (i,j) we have to determine $t_{i,j}$ and π_{P_j}. Instead of π_{P_j} we may use the probability of successfull follow–up from wave 1 to wave t which is smaller than π_{P_j}. This probability is given by $\pi_{j,1,t}/\pi_{j,1}$. Hence:

[7] One may compute the number of (i,j) pairs, where person i and person j live together in wave 1 and participate in the panel until wave 8. 71.3% of these pairs still live together in a household.

$$\hat{V}_{LL} = \sum_{i \in S}(\frac{1}{\pi_{i,1,t}} - 1)\frac{Y_i^2}{\pi_{i,1,t}} + \sum_{i \in S}\sum_{j \in H_{i,1} \cap S}(\frac{1}{\pi_{j,1,t}} - \frac{\pi_{j,1}}{\pi_{j,1,t}})\frac{Y_i Y_j}{\pi_{i,1,t}} \quad (18)$$

is a lower bound for $\hat{V}(\hat{T}_Y)$ which is easy to calculate.

For longitudinal tabulations that start at a later wave $t_s > 1$ we have to combine the results for cross-sections and longitudinal analysis.

To obtain a lower bound we use a smaller set $D_{i,t} \cap S$ and a lower bound for

$$(\frac{\pi_{i,j,t_s,t}}{\pi_{i,t_s,t}\pi_{j,t_s,t}} - 1)/\pi_{i,j,t_s,t} \quad (19)$$

where $\pi_{i,t_s,t}$ is the inclusion probability of person i in the panel waves t_s to t and $\pi_{i,j,t_s,t}$ is the corresponding joint longitudinal inclusion probability. According to the cross-sectional argumentation, we may use H_{i,t_s}, the set of persons living together with person i at wave t_s, as an adequate subset of $D_{i,t} \cap S$. By the same reasoning like in the case $t_s = 1$ we get the lower bound:

$$\hat{V}_{LL} = \sum_{i \in S}(\frac{1}{\pi_{i,t_s,t}} - 1)\frac{Y_i^2}{\pi_{i,t_s,t}} + \sum_{i \in S}\sum_{j \in H_{i,t_s} \cap S}(\frac{1}{\pi_{j,t_s,t}} - \frac{\pi_{j,t_s}}{\pi_{j,t_s,t}})\frac{Y_i Y_j}{\pi_{i,t_s,t}} \quad (20)$$

An upper bound for $\hat{V}(\hat{T}_Y)$ is obtained, if we replace the set of persons with dependent choice by $D_{i,t} \cap S$ and use an upper bound for the covariance term in eq. (19). Hence $D_{i,t} \cap S$ is the set of all persons that belong to the root-household of person i and are in the longitudinal sample S. This gives the upper bound:

$$\hat{V}_U = \sum_{i \in S}(\frac{1}{\pi_{i,t_s,t}} - 1)\frac{Y_i^2}{\pi_{i,t_s,t}} + \sum_{i \in S}\sum_{j \in D_{i,t_s} \cap S}(\frac{1}{\pi_{j,t_s,t}} - 1)\frac{Y_i Y_j}{\pi_{i,t_s,t}} \quad (21)$$

3 The use of random group estimation in panels

The preceding section presented only lower and upper bounds for the variance of cross-sectional and longitudinal totals. The estimation of the variance of more complicated population values gets more involved. This is true even for rather simple estimates like $\hat{D}_Q$ and $\hat{D}_L$ in the introduction. Therefore we need a simple variance estimator that applies for all interesting population values P.

In the special case of the SOEP there is a third motivation to look for a different variance estimator. Because of the use of the systematic sampling at the starting wave of the panel, the choice of households in wave one is not independent and the inclusion probability for neighbouring households is 0. So the justification for the use of eqn. (2) to estimate $\hat{V}(\hat{T}_Y)$ is lacking.

The method of ramdom groups (cf. Wolter 1985, ch. 2) is especially easy to adapt to the panel character of the sample and the SOEP's use of systematic sampling. Other resampling methods like Bootstrap and Jackknife are not regarded here because the rationale of these methods is model-based, i.e. they need a model for the characteristics of interest; see Rao/Wu (1988) and Sitter (1992) for a recent use of these techniques. The approach here is design-oriented; cf. also Rendtel (1993b). Thus no attempts are made here to model correlations between the characteristics of interest.

The original method needs R independent and identical distributed replications of the sampling experiment . Every replication of this sampling experiment gives a population estimate. The dispersion of these R independent and identically - distributed estimates provides an estimate of the theoretical dispersion.

Let $\hat{P}_r$ $(r = 1, \ldots, R)$ an estimate of P on the basis of the r^th replication. From $\hat{P}_r$ one may compute:

$$\hat{\bar{P}} = \frac{1}{R} \sum_{r=1}^{R} \hat{P}_r \tag{22}$$

which is also a reasonable estimate of P. Under the assumption that the

replications are independent and identical distributed

$$\hat{\hat{V}} = \frac{1}{R(R-1)} \sum_{r=1}^{R} (\hat{P}_r - \hat{\hat{P}})^2 \tag{23}$$

is an unbiased estimate of $V(\hat{\hat{P}})$. If $\hat{P}$ is a linear estimator, then $\hat{\hat{P}} = \hat{P}$. So $V(\hat{\hat{P}})$ is a reasonable estimate of $V(\hat{P})$ too.

In the case of the SOEP there was no initial replication of the sampling design. The sample was set up as a result of a single sampling experiment. The basic idea is here to produce such a subdivision of the original sample, that each subsample may be regarded as a realization of the original sampling design with reduced sample size.

In this case the choice of the subsamples is not independent and $\hat{\hat{V}}$ will not be unbiased for $V(\hat{\hat{P}})$. If the distribution of the $\hat{P}_r$ is identical (but not necessarily independent) the bias of $\hat{\hat{V}}$ is (cf. Wolter 1985 p.33) is given by:

$$E(\hat{\hat{V}}) = V(\hat{\hat{P}}) - Cov(\hat{P}_1, \hat{P}_2) \tag{24}$$

As a rule the $\hat{P}_r$ will be negatively correlated, so there is the tendency to over–estimate the variance of $V(\hat{\hat{P}})$.

Each subsample (called random group) has to have the same sampling design like the original sample. In the case of a household panel this implies that each random group results in a household panel with a reduced number of starting households. This is achieved by dividing the wave–1–households into subsamples. All persons and households that relate to the same root household are attributed to the random group of the root household. This rule works fine as long there are no persons from different root households that move together. As previously noted this is expected to happen rather seldom and did not occur in the SOEP up to wave 8. As a result of this procedure each random group is a panel on its own with the original follow–up rules.

The subdivision of the first wave households should "reproduce" their sampling design. Wolter (1985, pp. 30) lists some rules how this may be achieved for different sampling designs. This is epecially easy for systematic sampling with random start and interval. The random start is reproduced by a random draw of an interger $r \in \{1, \ldots, R\}$. The interval rule is reproduced

by enlarging the original interval by a factor R. This yields that every R^{th} sample unit belongs to the r^{th} subsample. This reflects also the sequencing of the units. Finally ramdom group $r+1$ is obtained from units $2, 2+R, \cdots$. Random group $r+2$ starts with unit 3 ... and so on, in modulo R fashion.

If two stage sampling is applied, it is usually recommended do divide the primary sampling units (PSU's) into subgroups. The rationale behind this strategy is the assumption that the variation between PSU's is bigger than the variation of the secondary units (SSU's) within the PSU's. Under this assumption the subsampling of the PSU's yields a conservative variance estimation.

The third element of the SOEP's first wave sampling is regional stratification. On the average there were only 4 PSU's per stratum. This is much to low to reproduce systematic sampling within each stratum and preserve simultaneously the strata proportions. In order to tackle this problem four different ramdom group formations were used:

R1: Ignores the strata and replicates the systematic sampling on the PSU's.

R2: Collapsed 2 neighbouring stata and replicates the systematic sampling within the new strata.

R3: Preserved the original strata but cuts each PSU into 2 new PSU's. The first half and the second half of the original PSU constituted the new PSU's. Systematic sampling was reproduced within the original strata on the basis of the new PSU's.

R4: The systematic sampling was reproduced on the SSU's within each PSU. This preserves strata proportions.

Since R2 and R3 produce stata of average size 8 and since the average number of SSU's per PSU is also 8, R was assigned the value 8.

Each random group formation R1 to R4 produces a variance estimate on it's own. Hence $\hat{\bar{V}}_R$, the average of these estimates, is also a reasonable estimate for the variance. As Särndal et al. (1992, p. 430) mentioned, the averaging of variance estimates is more precise than the single variance estimates.

One potential drawback of the ramdom group method is the treatment of non-response. The estimates $\hat{P}_r$ use weights that base on an analysis

of the panel attrition of the entire sample. A rigorous treatment of the random–group method would need weights that base on an separate drop–out analysis within each random group, see Wolter 1985 p. 83. The restriction to a non–response model that is based on the entire sample may lead to an underestimation of $\hat{\hat{V}}$, since the variance due to the model estimation is ignored.

A separate drop–out anaysis in each random group is rather complicated in a panel survey, since the drop–out analysis has to be repeated for each stage of the follow–up separately [8]. In the SOEP for each follow–up step a model for re–contact and a model for response of re–contacted households was used. So for 8 waves one has to estimate $(8-1) \times 2 \times R = 14 \times R$ drop out models for each formation R1 to R4. If one wants to present a methodological rigorous estimation of the variance by the average of R1 to R4, one has to estimate $14 \times 8 \times 4 = 448$ drop out models.

Such an analysis was performed by Rendtel (1993c) under some simplifiing restrictions. The results revealed:

- Almost no bias for the estimation of $\hat{\hat{V}}_{\bar{R}}$.
- Only moderate biases for the estimation of $\hat{\hat{V}}$ by the single random group formations R1 to R4.
- A high stability of $\hat{\hat{V}}_{\bar{R}}$ with respect to the use of drop–out models with differing covariates.

These empirical results are conform with other studies cited by Walter (1985, p. 84). In the following sections therefore only the simplified random group concept will be used.

4 The effect of panel attrition on the distibution of weights

In a household panel there are 3 sources that affect the distribution of weights:

[8] Also the analysis of inclusion probabilities of persons entering the panel (cf. Rendtel (1993b) has to be done separately.

(i) Increase of the variance of weights due to different inclusion probabilities for households with new persons.

(ii) Increase of the size of weights due to the reduction of the sample size by panel attrition.

(iii) Increase of the variance of weights due to differential non–response.

It has to be noted that source (i) is irrelevant for longitudinal tabulations that start from wave one. If the interval for tabulations starts at later waves, source (i) has to be accounted for like in the cross–sectional analysis.

Table 1 compares at the individual level sample sizes and some parameters of the empirical distribution of $W_{i,t} = 1/\pi_{i,t}$. Table 1 reveals that n_t, the cross–sectional sample size of the SOEP, dropped about 20% from wave 1 (12245 interviews) to wave 8 (9466 interviews). This drop in sample size is accompanied by an increase of N_t, the population size[9], of about 2 million persons. Hence $E(W_{i,t})$, the average of weighting factors, grew from 4089 to 5484. This is 1.39 times the wave 1 average. But also $CV(W_{i,t})$, the coefficient of variation grew from 0.816 to 0.977 due to sources (i) and (iii).

Figure 1 compares the distribution[10] of 3 weighting schemes:

- $W_{i,t} = \frac{1}{\pi_{i,1}}$ Cross–section wave 1
- $W_{i,t} = \frac{1}{\pi_{i,8}}$ Cross–section wave 8
- $W_{i,t} = \frac{1}{\pi_{i,1,8}}$ Longitudinal weight wave 1 to 8

The bimodal shape of the distributions in figure 1 clearly exhibits the different sampling probabilities for the two subsamples of the SOEP[11]. Furthermore one gets an impression how the panel attrition disperses the initial distribution of weights. This is true for the cross-sectional distribution but also for the longitudinal distribution.

[9]The population refers to all west–german persons older than 16 years living in private households.

[10]The density estimation in figure 1 was done by a kernel estimate, cf. Silverman (1986). The kernel function was a normal density function with standard deviation of $\sigma = 100$.

[11]In the foreigner subsample B the inclusion probabilities were 1/500 on the average, while in the west–german subsample A the average inclusion probability was approximately 1/5000.

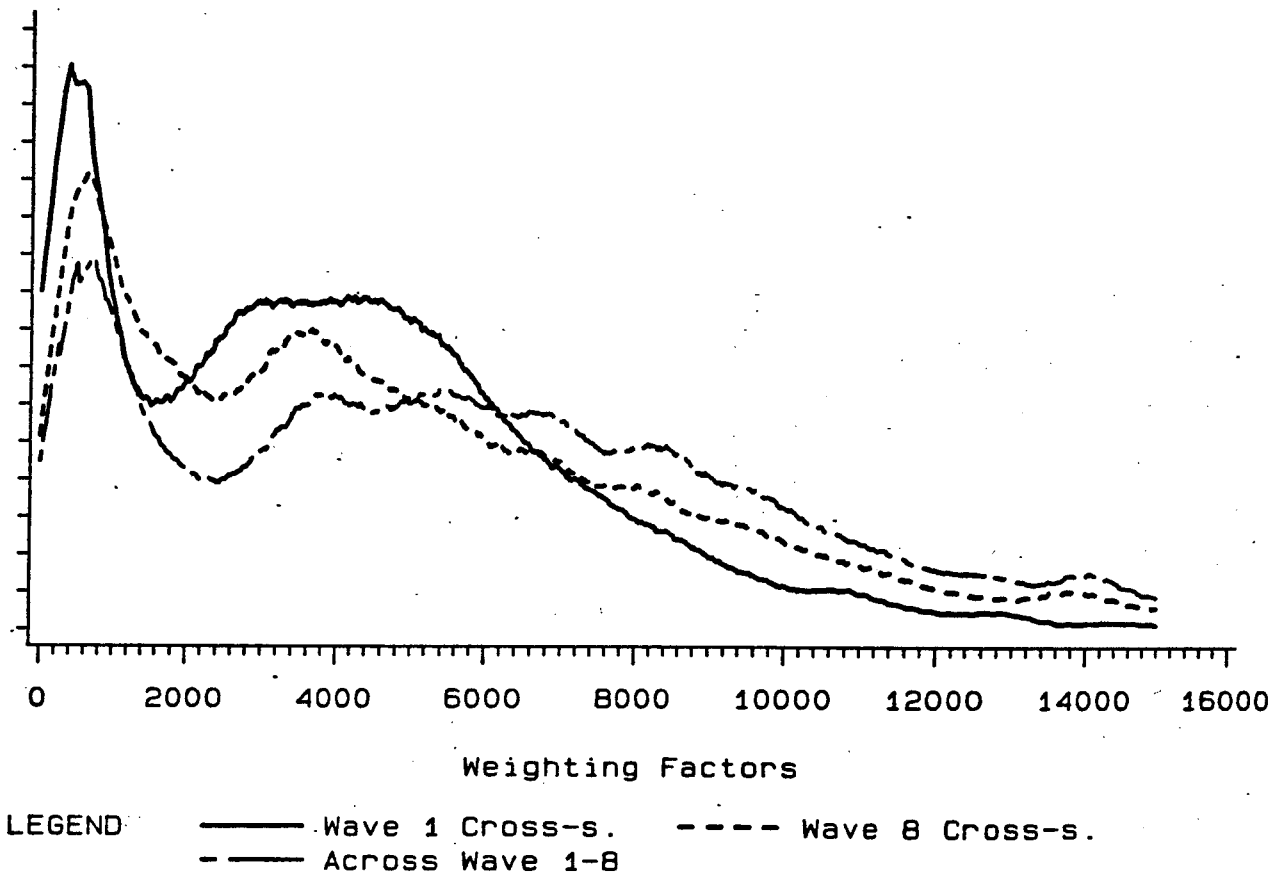

Figure 1: **The distribution of some SOEP cross–sectional and longitudinal weights**

Table 1: **Comparison of SOEP sample characteristics (persons in cross–sections): Sample size n_t, size of population N_t (in thousands), mean of weighting factors $E(W_{i,t})$ and its coefficient of variation $CV(W_{i,t})$ for waves $t = 1, \ldots, 8$.**

Wave	n_t	N_t	$E(W_{i,t})$	$CV(W_{i,t})$
1	12245	50073	4089	0.816
2	11090	49817	4492	0.823
3	10646	50216	4717	0.853
4	10516	50498	4802	0.874
5	10023	51227	5115	0.881
6	9710	51111	5263	0.919
7	9519	51499	5410	0.926
8	9467	51923	5484	0.977

5 A comparison of different variance estimators

In this section we want to compare the estimators for the variance of population totals that were introduced in the preceding sections. This comparison is done for 5 characteristics. The choice of these characteristics was due to the following considerations:

- "Household with child younger than 6" stands for a household characteristic.
- "Party preference for the Socialdemocrats" stands for an individual characteristic with high homogenity within households and possibly high stability with respect to time. The charactristic is known to be regionally concentrated. If systematic sampling offers gains in precision it should be indicated by this characteristic.
- "Fulltime employed woman" is a characteristic that induces no household cluster effects and is quiete stable over time.

- "Parttime employed" and "on vocational training" also induce no serious household cluster effects. They are believed to be less stable over time and less frequent than the above mentioned characteristics.

The variance estimates are displayed by the coefficient of variation, $CV(\hat{T}_Y)$, of $\hat{T}_Y$. The use of $CV(\hat{T}_Y)$ makes the variance of $\hat{T}_Y$ comparable for different characteristics. The following tables present $CV(\hat{T}_Y)$ for different variance estimators:

B_L, B_U: Lower and upper bounds for $\hat{V}(\hat{T}_Y)$.

$R1, R2, R3, R4$: Variance estimates by a single random group formation.

$\bar{R}$: Variance estimation by the mean of the variances from R1 to R4.

S: Variance estimation by the use of the formulas for simple sampling[12].

The tables also display the estimate $\hat{T}_Y$.

Table 2 compares the variance estimates for the wave 1 cross-sectional estimates. In this case the lower and upper bounds for $\hat{V}(\hat{T}_Y)$ coincide. The results may be summerized as follows:

- The approximation by simple sampling formulas is poor especially in cases of household cluster effects (party preference) and unequal sampling probabilities (vocational taining).

- The effect of systematic sampling appears to be ignorable, if one compares the estimates $B_L = B_U$ and $\bar{R}$ with respect to party preference. This appeares to be reasonable since the SOEP design used intensively regional stratification, which has approximately the same effect like systematic sampling. Both strategies produce a sample that spreads over the sampling area like a net.

- The random group estimates R1 to R4 exhibit a low stability. There is no apparent trend between R1 and R4. Such a trend would be resonable: R1 ignores the stratification of the PSU's. So it promises to give higher estimates of the variance. R4 measuers only the variation within the SSU's. In this case one expects lower variance estimates. R2 and R3 are in between the two extremes.

[12]The squared coefficient of variation is given by $\alpha(\frac{N}{\hat{T}_Y} - 1)$ with $\alpha = \frac{N-n}{(N-1)n}$.

Table 2: **Comparison of $CV(\hat{T}_Y)$ computed from different variance estimators. Population estimates for cross–sections in wave 1 of the SOEP.** Estimates of $\hat{T}_Y$ in thousands.

characteristic	$\hat{T}_Y$	S	$B_L = B_U$	R1	R2	R3	R4	$\bar{R}$
Household w. child younger than 6	3431	0.033	0.040	0.017	0.031	0.039	0.043	0.034
Party preference for Socialdemocrats	13145	0.015	0.029	0.040	0.014	0.028	0.018	0.027
Fulltime employed woman	5810	0.024	0.036	0.050	0.063	0.058	0.031	0.052
Partime employed	2658	0.038	0.048	0.036	0.026	0.046	0.036	0.037
On vocational training	1992	0.044	0.070	0.096	0.084	0.099	0.039	0.082

- The estimates $\bar{R}$ and $B_L = B_U$ are in good accordance for all five characteristics.

Table 3 gives the same comparison for wave 8 cross–sectional estimates. Again the S approximations are poor and the random group estimates R1 to R4 are not very precise allthough the range of the 4 variance estimates has decreased remarkably. The most interesting results of table 3 are:

- The differences between B_L andB_U are less than 10% for all characteristics.
- The estimates by $\bar{R}$ are in good accordance with the $[B_L, B_U]$ interval, although $\bar{R}$ not always lies within this interval[13].

Table 4 gives a comparison for longitudinal tabulations. There are two periods of equal length. Period 1 ranges from wave 1 to wave 4. Period 2

[13]Since $\bar{R}$ oparates on a different estimation method and the interval is small, one may not expect the $\bar{R}$–estimate to fall into the $[B_L, B_U]$–interval.

Table 3: **Comparison of $CV(\hat{T}_Y)$ computed from different variance estimators. Population estimates for cross-sections in wave 8 of the SOEP.** Estimates of $\hat{T}_Y$ in thousands.

Characteristic	$\hat{T}_Y$	S	B_L	B_U	R1	R2	R3	R4	$\bar{R}$
Household with child younger than 6	3634	0.038	0.053	0.057	0.032	0.060	0.061	0.061	0.055
Party preference for Socialdemocrats	12834	0.018	0.036	0.038	0.035	0.032	0.031	0.033	0.033
Fulltime employed woman	6679	0.027	0.041	0.043	0.049	0.046	0.021	0.028	0.037
Parttime employed	4195	0.035	0.049	0.050	0.035	0.041	0.044	0.045	0.041
on vocational Training	1671	0.057	0.095	0.095	0.084	0.089	0.070	0.075	0.080

lasts from wave 5 to wave 8. The comparison is performed for 3 individual[14] characteristics, which check the stability of cross–sectional characteristics over time. The vocational training has been omitted here, because the regular training time in Germany is shorter than the period regared here.

The results of table 4 reveal for both periods:

- The poor approximation of the simple sampling formulas.
- A high variabiliy of the single random group estimates R1 to R4.
- Ignorable differences between the variance bounds[15].
- A good accordance of the $\bar{R}$–estimate with the variance bounds.

These results confirm the cross–sectional findings. As a result of this section we found that the variance estimation by $\bar{R}$ produced good approximations for the variance of estimates of populations totals. This may serve as an justification to use the variance estimation by $\bar{R}$ in cases where there are no such simple variance bounds.

6 The inflation of the variance of estimates for population totals

In this section we want analyze the inflation of the variance of estimates for population totals. In order to study the effect of the panel attrition we refer to a fictious experiment. Suppose that we draw at wave t a sample with exactly the same sampling design as in wave 1. Such a fictious sample would have the same variance properties like the wave–1–sample of the panel. Our aim is to compare the variance of the wave–t panel sample with the variance that would be obtained by an estimate from the fictious sample.

Such a comparion is confronted with the problem that the variance of an estimate of a population characteristic depends not only on the design of the

[14]Since households are not stable longitudinal units there is no longitudinal household unit here. See Ernst (1989) for a discussion of "longitudinal" households.

[15]The differences were smaller than the last digit that is displayed in the tables for the coefficient of variation. Hence in table 4 the results for $B_L L$, B_L and B_U are presented in a joint column.

Table 4: **Comparison of $CV(\hat{T}_Y)$ computed from different variance estimators. Longitudinal estimates for two periods of the SOEP.** Period 1 : 1984 until 1987 (waves 1 to 4). Period 2 : 1988 until 1991 (waves 5 to 8). Estimates for $\hat{T}_Y$ in thousands.

Characteristic	$\hat{T}_Y$	S	$B_L \approx B_U$	R1	R2	R3	R4	$\bar{R}$
	Period 1: waves 1 to 4							
Continouus party preference for Socialdemocrats	6897	0.025	0.045	0.053	0.022	0.056	0.015	0.041
All waves fulltime employed woman	4039	0.035	0.048	0.060	0.057	0.045	0.037	0.051
All waves part-time employed	1263	0.064	0.077	0.057	0.061	0.072	0.054	0.061
	Period 2: waves 5 to 8							
continouus party preference for Socialdemocrats	8064	0.025	0.045	0.041	0.035	0.044	0.045	0.042
All waves fulltime employed woman	4104	0.036	0.052	0.049	0.065	0.035	0.030	0.047
All waves part-time employed	1556	0.061	0.074	0.047	0.076	0.028	0.080	0.062

sample but also on the variance of the characteristic in the population. In the case of simple sampling the squared coefficient of variation, i.e. $\hat{V}(\hat{T}_Y)/(T_Y)^2$, is proportional to $1/\bar{P} - 1$, where $\bar{P} = T_Y/N$. In order to disentangle the effects of panel attrition and frequency in the population we assume that $CV_f(\hat{T}_{Y_t})$, the coefficient of variation for $\hat{T}_{Y_t}$ in the fictious sample in wave t, behaves in the same way. Thus

$$\frac{CV_f^2(\hat{T}_{Y_t})}{1/\bar{P}_t - 1} = \frac{CV^2(\hat{T}_{Y_1})}{1/\bar{P}_1 - 1} \tag{25}$$

for $t = 2, 3, \ldots$ .Note that this assumption is much weaker than the strict relationship of $CV(\hat{T}_Y)$ and $\bar{P}$ in the case of simple sampling. In the latter case the ratio of $CV(\hat{T}_Y)$ and $\sqrt{1/\bar{P} - 1}$ is the same for all characteristics while here this ratio may differ from one characteristic to another[16]. Note also that it is hard to check such an assumption by empirical evidence. Under this assumption the coefficient of variation of $\hat{T}_{Y_t}$ under the fictious sample is given by:

$$CV_f(\hat{T}_{Y_t}) = CV(\hat{T}_{Y_1})\sqrt{\frac{1/\bar{P}_t - 1}{1/\bar{P}_1 - 1}} \tag{26}$$

Table 5 compares η_t, the ratio of $CV(\hat{T}_{Y_t})$ from the panel and $CV_f(\hat{T}_{Y_t})$ from the fictious sample, for the first 8 waves of the SOEP. It also displays $\bar{P}_t$, the estimated proportion of the characteristic in the wave t population. The lower bound estimate B_L, which appeared to be more stable than the $\bar{R}$–estimate,was used for the variance estimation in table 5. The differences between B_L and B_U are so small that they are ignored here.

Table 5 reveals a more or less steady increase of η_t. Since the variance estimates are also subject to some random fluctuation one may not expect a perfect monotone increase of η_t. This apperars to be relevant for the characteristic "On vocational training", where the proportion in the population is

[16]One may compute this ratio for the five characteristics in wave 1; see table 2. If one uses the variance estimates by bound B_L from table 2 and the population percentages displayed in table 5 (see below) one gets the ratios: 0.015, 0.017, 0.013, 0.011 and 0.014. This reveals a considerable variability of the above mentioned ratio over characteristics.

Table 5: **The effect of panel attrition on the standard deviation of $\hat{T}_{Y_t}$. Cross-sectional estimates from the SOEP.** $\bar{P}_t$ = estimated proportion in the populaton (in Percent). η_t = increase of $\hat{V}^{1/2}(\hat{T}_{Y_t})$ with respect to wave 1

Wave	Household with child younger 6 y.		Party preference for Socialdemocrats		Fulltime employed woman		Parttime employed		On vocational training	
	$\bar{P}_t$	η_t	$\bar{P}_t$	η_t	$\bar{P}_t$	η_t	$\bar{P}_t$	η_t	$\bar{P}_t$	η_t
1	13.0	1.00	26.3	1.00	11.6	1.00	5.3	1.00	4.0	1.00
2	10.8	0.99	25.9	1.06	12.3	1.12	6.5	1.03	2.7	0.99
3	11.1	0.92	25.9	1.06	12.6	1.12	5.9	1.12	2.9	0.90
4	13.6	1.18	25.6	1.08	12.5	1.16	6.5	1.12	3.8	0.95
5	11.5	1.09	26.9	1.12	12.4	1.16	6.6	1.13	3.9	0.97
6	11.9	1.15	27.0	1.16	13.1	1.19	6.7	1.18	3.5	1.03
7	12.2	1.18	28.3	1.17	12.6	1.19	7.5	1.25	3.3	1.02
8	12.5	1.27	24.7	1.16	12.8	1.21	8.1	1.28	3.2	1.21

Table 6: **The effect of panel attrition on the standard deviation of $\hat{T}_{Y_t}$. Longitudinal estimates for different periods from the SOEP.** $\bar{P}_t$ = estimated proportion in the populaton (in Percent). η_t = increase of $\hat{V}^{1/2}(\hat{T}_{Y_t})$ with respect to first period.

	Continouus party preference for Socialdemocrats		All waves fulltime employed woman		All waves parttime employed	
Period	$\bar{P}_t$	η_t	$\bar{P}_t$	η_t	$\bar{P}_t$	η_t
1984-1987	14.5	1.00	8.5	1.00	2.6	1.00
1985-1988	15.8	1.05	8.5	1.04	2.7	0.99
1986-1989	16.3	1.07	8.6	1.07	2.7	1.03
1987-1990	17.4	1.05	8.5	1.04	3.2	1.07
1988-1991	16.6	1.08	8.4	1.08	3.2	1.06

rather small (about 3.5%) and as a consequence also the variance estimate is not very precise.

As table 5 indicates, the variation of the population proportions $\bar{P}_t$ are of moderate size, so the implications of assumption (26) will not be substantial. Until wave 8 there is a non-ignorable loss in the precision of estimates of population totals, which varies with respect to the 5 characteristics regarded here.

A similar analysis was performed for the longitudinal characteristics of table 4. Here the 4-year period was subsequently shifted from wave 1 to 4 (1984 - 1997) by one year until the period wave 5 to 8 (1988 - 1991). Table 6 also uses the B_L-variance estimate.

Also in the longitudinal case there appears a notable loss in the precision of estimates, although the losses seem to be somewhat smaller than is the corresponding cross-sectional cases.

As mentioned above, the inflation of the variance has two sources: The reduction of the sample size and the increase in the variance of weights. The proportion due to the reduction of the sample size may be estimated by the ratio:

$$\kappa_t = \sqrt{\frac{N_t/n_t}{N_1/n_1}} \tag{27}$$

where N_t is the population size and n_t is the sample size at wave t[17]. κ_t denotes the increase of the standard deviation of $\hat{T}_{Y_t}$ if all inclusion probabilities are uniformly increased by a constant factor. This is seen by replacing all probabilities π_i in eqn. (1) by π_i/κ.[18]

In the case of the SOEP for t=8 and cross-section we have $\kappa_8 = 1.16$ (individual level) and $\kappa_8 = 1.18$ (household level). For longitudinal tabulation we have $\kappa_t = 1.06$, if we compare sample and (estimated) population size in period 1984 - 1988 with period 1989 - 1991. For the cross-sectional characteristic "party preference" and all longitudinal characteristics the variance inflation mets exactly the value of κ. But also in the case of the other characteristics the variance inflation is only slightly above this bound. The biggest increase appears for the characteristic "parttime employed". Here the inflation factor is $\tilde{\kappa}_8 = 1.28$. We may calculate the value $\tilde{n}_t = fn_t$ that causes such an inflation $\tilde{\kappa}_8 = 1.28$ by a further reduction of the sample size. By eq. (27) we get:

$$f = (\frac{\kappa_t}{\tilde{\kappa}_t})^2 \tag{28}$$

which yields f=0.95 in the case of the characteristic "parttime employed". Thus the effect of differential non-response (and different inclusion probabilities with respect to move-in's) is small as compared to the reduction of the sample size. The additional variance effect of differential non-response until wave 8 is equivalent to a uniform reduction of the sample size by a factor f=0.95. This is a fairly small percentage compared with the decrease in sample size from wave 1 to wave 8, which is 0.773. Thus we may conclude that the most important variance effect of panel attrition is the reduction of the sample size.

[17]This notation includes also the longitudinal case, where N_t denotes the longitudinal population from wave 1 to wave t and n_t the corresponding sample size.

[18]Indeed replacing $\frac{\kappa}{\pi_i} - 1$ by $\frac{\kappa}{\pi_i} - \kappa$ yields some underestimate of the variance effect by (27), which will be small since $1/\pi_i \gg 1$.

Although there may be characteristics with greater effects of differential non–response, the above result is a plausible consequence of the drop–out analysis of the SOEP, see Rendtel (1990,1993a,1993c). The result of this analysis may be condensed in the paradigma that drop–out is a matter of fieldwork and not a matter of socio–economic characteristics. As long as the most important variables for fieldwork — like change of the interviewer, split–off of a household or item non–response for household income — are not correlated with characteristics of interest, the panel attrition results simply in a mere reduction of sample size.

7 The precision of different trend estimators

In this section we want to compare the precision of the trend estimators

$$\hat{D}_C = \sum_{S_{t_2}} \frac{1}{\pi_{i,t_2}} Y_{i,t_2} - \sum_{S_{t_1}} \frac{1}{\pi_{i,t_1}} Y_{i,t_1} \tag{29}$$

and

$$\hat{D}_L = \sum_{S_{t_1} \cap S_{t_2}} \frac{1}{\pi_{i,t_1,t_2}} d_i + \sum_{S_{t_2} \setminus G_{t_1}} \frac{1}{\pi_{i,t_2}} Y_{i,t_2} - \sum_{S_{t_1} \setminus G_{t_2}} \frac{1}{\pi_{i,t_1}} Y_{i,t_1} \tag{30}$$

of the population value $D = T_{Y_2} - T_{Y_1}$.

Note that in the special case of a household panel $\hat{D}_C$ and $\hat{D}_L$ are not equivalent even if there is no panel attrition. The difference results from persons that have entered the panel sample by a move into a panel household. Thus — assuming on panel attrition — we have $\pi_{i,t_1} = \pi_{i,t_1,t_2}$ but $\pi_{i,t_1} \neq \pi_{i,t_2}$.

The estimate $\hat{D}_L$ does not use the sample information from persons that move into panel households, unless they are entering also the population. The estimator $\hat{D}_C$ exploits the information of <u>all</u> sampled persons in wave t_1 and wave t_2 but it doesn't exploit the information efficiently since it does not base on the individual differences $d_i = Y_{i,t_2} - Y_{i,t_1}$.

In this section we use the variance estimation by $\bar{R}$, since there are no "simple" bounds available like in the case of the population totals.

Table 7: **Comparison of the precision of $\hat{D}_C$ and $\hat{D}_L$. Estimated population value $D = Y_{1991} - Y_{1984}$. Estimates from the SOEP wave 1 (1984) and wave 8 (1991).** Estimates for population values in thousands. $\Delta = \sigma(\hat{D}_L)/\sigma(\hat{D}_C)$

	Party preference for Socialdemocrats	Fulltime employed woman	Parttime employed	On vocational training
$\hat{D}_C$	-311	869	1537	-321
$CV(\hat{D}_C)$	1.537	0.329	0.129	0.633
$\hat{D}_L$	-1104	309	1301	-319
$CV(\hat{D}_L)$	0.359	0.760	0.143	0.751
	Ratio of standard deviations for $\hat{D}$			
Δ	0.83	0.82	0.94	1.18

Table 7 displays the comparison of the two estimates and their variances for the individual[19] characteristics of tables 2 and 3. The time interval was t_1 = wave 1 and t_2 = wave 8.

The direction of the estimated change coincides for 2 estimators, but there are remarkable differences in the size of the estimated trend. For example, the estimated decrease of the number of persons with a preference for the Socialdemocrats by $\hat{D}_L$ is 3 times bigger than the corresponding estimate by $\hat{D}_C$. Since the $2 \times \sigma$ confidence intervals of both estimates overlap the two estimates do not exclude each other.

If we compare the ratio Δ of the standard errors of $\hat{D}_L$ and $\hat{D}_C$ we see that there may be a remarkable gain in precision by the use of $\hat{D}_L$. The gain in precision depends on the stability of the characteristic with respect to time.

[19]The household characteristic was omitted here since $\hat{D}_L$ needs a longitudinal identification of a sample unit, which is difficult for households.

8 The loss of precision of trend estimates

In this section we want to balance the benefits of the panel design for trend estimates against the losses in precision from panel attrition.

Again we refer to a fictious independent replication of the wave–1 sampling design. This fictious wave–t_2 sample gives a population estimate $\hat{T}_{f,Y_{t_2}}$, which may be used for an alternative trend estimate:

$$\hat{D}_f = \hat{T}_{f,Y_{t_2}} - \hat{T}_{Y_{t_1}} \tag{31}$$

Since $\hat{T}_{f,Y_{t_2}}$ and $\hat{T}_{Y_{t_1}}$ are independent the variance of $\hat{D}_f$ is computed as:

$$V(\hat{D}_f) = V(\hat{T}_{f,Y_{t_2}}) + V(\hat{T}_{Y_{t_1}}) \tag{32}$$

The use of eqn. (25) yields:

$$V(\hat{D}_f) = V(\hat{T}_{Y_{t_1}})(1 + \frac{\hat{T}^2_{Y_{t_2}}(\frac{1}{\bar{P}_{t_2}} - 1)}{\hat{T}^2_{Y_{t_1}}(\frac{1}{\bar{P}_{t_1}} - 1)}) \tag{33}$$

Thus the balance of the gains from the panel design and the losses from the panel attrition may be expressed by the ratio:

$$\gamma = \frac{V^{1/2}(\hat{D}_{\text{Panel}})}{V^{1/2}(\hat{D}_f)} \tag{34}$$

Table 8 displays this ratio for $\hat{D}_{\text{Panel}} = \hat{D}_C$ and $\hat{D}_{\text{Panel}} = \hat{D}_L$. The variance of $\hat{T}_{Y_{t_1}}$, $\hat{D}_C$ and $\hat{D}_L$ was estimated by the $\bar{R}$ random group estimator.

Table 8 reveals for some cases still a remarkable superiority of the panel benefits after 8 panel waves. This is true for stable characteristics like "party preference" and "fulltime employed woman". In case of instable characteristics like "parttime employed" and "on vocational training" the panel design offers no benefits[20], so losses are expected to dominate. But even in this case the losses from panel attrition are still moderate, while the gains by the panel designs — in cases of stable characteristics — appear to be substantial.

Besides, one must not forget that a sequence of independent cross–sectional surveys is no capable to produce longitudinal tabulations, which is the main focus of panel surveys.

[20]In such cases the covariance $C(\hat{T}_{Y_{t_2}}, \hat{T}_{Y_{t_1}})$ is almost zero due to the instability of the characteristic.

Table 8: **Comparision of the ratio γ of the standard deviation of $\hat{D}_C$ and $\hat{D}_L$ with the $\hat{D}_I$ from a hypothetical trend estimate. Data from the SOEP wave 1 (1984) and wave 8 (1991).** Estimated trend: $D = T_{Y_{1991}} - T_{Y_{1984}}$

Characteristic	Value of γ for $\hat{D}_C$	$\hat{D}_L$
Party preference for Socialdemocrats	0.95	0.78
Fulltime employed woman	0.64	0.52
Parttime employed	1.26	1.18
On vocational training	0.91	1.07

A The sampling design of the SOEP

This appendix summarizes the sampling procedures of the SOEP. A detailed description may be found in Wagner et al. (1993).

Subsamples: The households of the first wave of the SOEP were taken from two subsamples[21]:

A: Households with german[22] head.

B: Households with foreign head. The nationality of the head had to be turk, jugoslav, greec, italian or spanish.

The foreigner households are oversampled.

2–stage sampling: For each subsample 2–stage sampling was used. The PSU's were in subsample

[21]There exists a third subsample C (East-Germans), which was started in 1990. This subsample is not referred here, since the attrition analysis presented here bases entirely on the subsamples started in 1984.

[22]To be precise: Nationality different from the nationalities sampled in subsample B.

A: Electorial districts

B: Regional districts or towns

The SSU's of the sample were in subsample

A: Households of the electorial district

B: Persons older than 16 in the foreigner register of the district/town

Systematic sampling: For each stage systematic sampling by interval and random start was used. The sequencing and the size for the PSU's was due to :

A: Regional sequencing and size proportinal to the estimated number of households in the PSU.

B: Regional sequencing of PSU's and size proportinal to the number of persons with the specific nationality in the PSU.

The sequencing and size for SSU's was due to:

A: Deterministic traverse rules for the PSU starting from a random adress ("Random Route"). Each 7^{th} household bell along the traverse was selected.

B: Sequencing according to the foreigner register. Size proportional to the number of household members in the register older than 16.

Stratification: The selection ot the PSU's of subsample A was done in two phases:

Phase I: The phase I ended in the 1982 ADM master sample (cf. Kirschner (1984) for a detailed description of the ADM sampling design). The sample was taken according to the above mentioned systematic sampling rules.

Phase II: The part of the master sample that was allocated to the fieldwork institute was stratified according to regional criteria. The strata sizes were proportional to the estimated number of households of the stata. Within each stratum there was systematic sampling according to the above rules.

Selection of persons: All persons older than 16 that live in a selected household are asked for an interview. In wave 1 only households with complete participation were sampled. In later waves this rule was ignored.

Follow–up rules: All interviewed persons will be followed if they move within the sampling area (until 1990: West–Germany, later: Germany). People that refuse to participate will no be followed; also people with two successive non–responses.

References

Ernst, L. (1989): *Weighting Issues for Longitudinal Household and Family Estimates.* In:Kasprzyk, D.; Duncan, G.; Kalton, G.; Sing, M. (Eds.) (1989): *Panel Surveys.* Wiley, New York, p. 139–159.

Frick, J.; Rendtel, U.; Wagner, G. (1993): *Eine Strategie zur Kontrolle von Längsschnittgewichtungen am Beispiel des Sozio–oekonomischen Panel (SOEP).* DIW–Discussion Paper No. 80, Berlin.

Hanefeld, U. (1984): *The German Socio-Economic Panel.* In: American Statistical Association (Ed:): 1984 Proceedings of Social Statistics Section, p. 117–124, Washington D.C.

Hill, M. (1992): *The Panel Study of Income Dynamics; A User's Guide.* Sage Publications, Newburg Park.

Kirschner, H.–P. (1984): *ALLBUS 1980: Stichprobenplan und Gewichtung.* In: Mayer, K.–U.; Schmidt, P. (Hrsg): Allgemeine Bevölkerungsumfrage Sozialwissenschaften, Frankfurt/M., 114–182.

Rao, J. N. K.; Wu, C. F. J. (1988): *Resampling Inferences with Complex Survey Data.* Journal of the American Statistical Association, 83, 231–241.

Rendtel, U. (1990): *Teilnahmebereitschaft in Panelstudien: Zwischen Beinflussung, Vertrauen und Sozialer Selektion.* Kölner Zeitschrift für Soziologie und Sozialpsychologie, 42, 280–299.

Rendtel, U. (1091): *Die Schätzung von Populationswerten in Panelerhebungen.* Allgemeines Statistisches Archiv, 42, 280–299.

Rendtel, U. (1993a): *Über die Repräsentativität von Panelstichproben. Eine Analyse der feldbedingten Ausfälle im Sozio–oekonomischen Panel (SOEP).* DIW–Discussion Paper No. 70, Berlin.

Rendtel, U. (1993b): *Design–oriented Weighting of a Household Panel.* DIW–Discussion Paper No. 79, Berlin.

Rendtel, U. (1993c): *Die Auswertung von Paneldaten unter Berücksichtigung von Panelmortalität.* Unpublished Manuscript.

Särndal, C.-E.; Swenssòn, B.; Wretman, J. (1992):*Model Assisted Survey Sampling.* Springer-Verlag, New York.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis.* Chapman and Hall; London.

Sitter, R. R. (1992): *A Resampling Procedure for Complex survey Data.* Journal of the American Statistical Association, 87, 755–765.

Wagner, G.; Burkhauser, R.; Behringer, F. (1993):*The English Language Public Use File of the German Socio-Economic Panel.* Journal of Human Resources, 28, p. 429–433.

Wagner, G.; Schupp, J.; Rendtel, U. (1993): *Das Sozio-oekonomische Panel — Methoden der Datenproduktion und -aufbereitung im Längsschnitt.* In: Hauser et al. (Hrsg): Mikroanalytische Grundlagen der Gesellschaftspolitik — Erhebungsverfahren, Analysemethoden und Mikrosimulation; Akademie Verlag, Berlin.

Wolter, K.M. (1985): *Introduction to Variance Estimation.* Springer, New York.