OPTIMIZING RESPONSE BURDEN IN PANELS

Adriaan Hoogendoorn *)
Dirk Sikkel **)

ABSTRACT

When maintaining a consumer panel one is sometimes confronted with the question of how large the response burden can be. It is our experience that the larger the response burden (i.e. the number and size of the respondent tasks), the higher the panel attrition. Consequently, we have to solve a trade-off problem. In this paper we derive some rules for optimum respondent burden for change estimators under simple model assumptions. The consequences of these results for daily research design in frequently measuring panels are discussed.

*) Nederlands Instituut voor Markt- en Maatschappijonderzoek (NIMMO)
**) University of Amsterdam, Faculty of Political and Socio-Cultural Sciences, Department of Methodology

The authors thank Willem Saris and Frank van de Pol for support and comments on earlier drafts of this paper

OPTIMIZING RESPONSE BURDEN IN PANELS

1. Introduction

In budget surveys and consumer panels one has to deal with problems concerning response burden. The work load associated with filling out diaries is considered to be a major reason for the high initial non response rate for budget surveys compared to other household surveys (Lindström, 1989; Lyberg, 1991). An other effect of response burden is found in underreporting. It is common to find that the number of reported purchased products in the first week is higher than in the second week (Harrison, 1991; Nevraumont, 1991; Ribe, 1991). In consumer panels the response burden influences the panel attrition (Silberstein & Jacobs, 1989). Modern techniques, like bar scanning methods and electronic diaries (Saris et al, 1992) are applied to relieve the respondent's task. But in spite of these techniques a respondent still has to do a considerable amount of work, especially in panels that aim at continuous measurement of expenditures. It may be, however, unnecessary to collect the budget data for every week. When we take only a sample of weeks this may result in a relative small loss of precision. On the other hand, it may result in a lower attrition rate as less respondents become fed up with their task.

In this paper we study the relation between response burden and the precision of estimators of change in consumer panels. Response burden is defined here as the number of weeks during a certain period (e.g. three months or a year). Although it affects initial non response and underreporting as well, we will focus on the effect on panel attrition. Therefore we assume a simple model which describes the attrition as a function of the response burden. In exploring the effect of panel attrition to the precision of our estimators, we will concentrate on the effect to the variance, not the bias. Most literature on panel attrition deals with bias effects. Several methods have been developed to deal with attrition bias. Examples are the use of Markov-chain models for nonrandom non response to estimate gross flows in categorical data (Stasny, 1987), econometric regression analyses to correct for attrition bias (Hausman and Wise, 1979), bias reduction by sample designs (Van de Pol, 1989) and bias corrections by weighting techniques (Van de Pol, 1993). We will assume that the households that dropped out of the panel are

immediately replaced using a quota sampling method (Van de Pol, 1989). Consequently the bias due to attrition is kept to a minimum, and the panel size is constant over time.

We will study two models that correspond with two panel designs. Model 1 deals with a panel that is dedicated to measure expenditures on consumption goods. This is the case in most consumer panels. In model 2, however, the panel is also used for other purposes. An example of this is the Dutch Stichting Telepanel, where the burdensome questionnaires on expenditures are interchanged with questionnaire on a variety of other topics. The theory derived here is stated in general terms as to make it applicable to a more general situation.

2. Preliminary notation and relations

We are interested in estimates (e.g. of consumption or purchases of fast moving consumer goods) over a certain period of M weeks. Usually M is equal to 13, a three month period, but we also may consider M=4, M=26 or M=52. In every wave (or week) we have n households from a much larger population of N households. It is assumed that there is a constant attrition p. For each week every panel member has a probability p to drop out of the panel independent of the other panel members and independent of what happens in other weeks. Hence, a respondent who is in the panel at week j has probability p^{k-j} to still be a panel member at week k>j. Let

$X_i^{(t)}_j$ be the amount of purchases by household i in week j of period t.

$X_j^{(t)} = Nn^{-1} \sum_{i=1}^{n} X_i^{(t)}_j$, the estimated population total for week j of t.

$X^{(t)} = \sum_{j=1}^{M} X_j^{(t)}$ the estimated population total for period t

It is assumed that for given j and t the $X_i^{(t)}_j$ are i.i.d. with variance σ^2 for the households i. For different values of j and t we assume the $X_i^{(t)}_j$ to be homoscedastic (possessing equal variances). We focus our attention to more or less daily shopping routines. We assume that this is a stable process which is in equilibrium, although it may be different for each household. When there

is a regular weekly pattern in such purchases it is reasonable to assume that for a combination $(j,t) \neq (k,u)$ we have

$$\mathrm{cov}(X_{ij}^{(t)}, X_{ik}^{(u)}) = \rho\sigma^2$$

Of course, such an assumption is violated for a product which does not follow such a weekly pattern, e.g. when it is purchased on a two-weekly basis. Such types of variables are outside the scope of this paper. This may suggest that the assumptions are rather restrictive. When, however, broad categories are used like meat, green vegetables, fruit or candies, the assumptions apply, at least in the Dutch society, to the most important results that have to come out of a budget survey.

Our main interest is to measure the changes from one period to another. The absolute consumption level of a product in itself is not a very useful figure. In terms of the variables defined above this means that we are interested in the precision of $X^{(t+1)}-X^{(t)}$ (and, as a byproduct, of $X^{(t)}$). Consequently, we require the covariances of the terms of which these quantities consist. Let n_{jk} be the number of panel members which are in both wave j and in wave k>j in a period t. Then n_{jk} has a binomial distribution with parameters n and p^{k-j}. Assume that the panel members are numbered such that the first n_{jk} are in both waves. Then we can compute

$$\begin{aligned}
\mathrm{cov}\,(X_j^{(t)}, X_k^{(t)}) &= E_{n_{jk}} \mathrm{cov}(X_j^{(t)}, X_k^{(t)} | n_{jk}) + \mathrm{cov}_{n_{jk}} (EX_j^{(t)}, EX_k^{(t)} | n_{jk}) \\
&= E\, \frac{N^2}{n^2} \sum_{i=1}^{n_{jk}} \mathrm{cov}(X_{ij}^{(t)}, X_{ik}^{(t)}) \\
&= N^2 p^{k-j} \rho\sigma^2 / n \qquad (1)
\end{aligned}$$

This yields the following variance for $X^{(t)}$:

$$\text{var}(X^{(t)}) = \text{var}\left(\sum_{j=1}^{M} X_j^{(t)}\right)$$

$$= \sum_{j=1}^{M}\sum_{k=1}^{M} \text{cov}(X_j^{(t)}, X_k^{(t)})$$

$$= \frac{N^2\sigma^2}{n}\left[M + \frac{2\rho(M-1)p}{1-p} - \frac{2\rho p^2(1-p^{M-1})}{(1-p)^2}\right] \qquad (2)$$

and for $X^{(t+1)}-X^{(t)}$:

$$\text{var}(X^{(t+1)}-X^{(t)}) = \text{var}\left(\sum_{j=1}^{M} X_j^{(t+1)} - \sum_{j=1}^{M} X_j^{(t)}\right)$$

$$= \text{var}(X^{(t+1)}) + \text{var}(X^{(t)}) - 2\sum_{j=1}^{M}\sum_{k=1}^{M} \text{cov}(X_j^{(t+1)}, X_k^{(t)})$$

$$= \frac{2N^2\sigma^2}{n}\left[M + \frac{2\rho(M-1)p}{1-p} - \frac{2\rho p^2(1-p^{M-1})}{(1-p)^2} - \frac{\rho p(1-p^M)^2}{(1-p)^2}\right] \qquad (3)$$

These relations hold when the respondents are required to fill in the questionnaire during all M weeks of periods t and t+1. When the respondents are required to fill in the questionnaire during only m out of M weeks we can distinguish two different models, corresponding with two different panel designs. In model 1 we assume that immediately after the measurement a fraction q=1-p to drops out of the panel. In the weeks when no measurement with respect to the $X_{ij}^{(t)}$ takes place there is no attrition. This corresponds to a single purpose panel, in which in each wave the household budgets are measured. In model 2, in contrast, the panel attrition continues in the weeks when no measurement with respect to the $X_{ij}^{(t)}$ takes place. This is the typical case of a telepanel, which may be used for many purposes. Both models lead to slightly generalized versions of (2) and (3). We generalize our definition of $X_j^{(t)}$ to

$$X_j^{(t)} = Mm^{-1}Nn^{-1}\sum_{i=1}^{n} X_{ij}^{(t)} \qquad (4)$$

So that $X^{(t)} = Mm^{-1}Nn^{-1}\Sigma_i \Sigma_j X_i^{\{t\}}$ where j takes the values of only those weeks in which respondent i fills in a questionnaire. Then for model 1 we have

$$\text{var}(X^{(t)}) = \frac{M^2N^2\sigma^2}{nm^2}\left(m + \frac{2\rho(m-1)p}{1-p} - \frac{2\rho p^2(1-p^{m-1})}{(1-p)^2}\right) \quad (5)$$

and

$$\text{var}(X^{(t+1)}-X^{(t)}) = \frac{2M^2N^2\sigma^2}{nm^2}\left(m + \frac{2\rho(m-1)p}{1-p} - \frac{2\rho p^2(1-p^{m-1})}{(1-p)^2} - \frac{\rho p(1-p^m)^2}{(1-p)^2}\right) \quad (6)$$

For model 2 we have

$$\text{var}(X^{(t)}) = \frac{MN^2\sigma^2}{nm}\left(M + \frac{2\rho(m-1)p}{1-p} - \frac{m-1}{M-1}\frac{2\rho p^2(1-p^{M-1})}{(1-p)^2}\right) \quad (7)$$

and

$$\text{var}(X^{(t+1)}-X^{(t)}) =$$

$$= \frac{2MN^2\sigma^2}{nm}\left(M + \frac{2\rho(m-1)p}{1-p} - \frac{m-1}{M-1}\frac{2\rho p^2(1-p^{M-1})}{(1-p)^2} - \frac{m-1}{M-1}\frac{\rho p(1-p^M)^2}{(1-p)^2}\right) \quad (8)$$

provided that the sampling design is balanced with respect to first and second order inclusions of the weeks in the sample, i.e. every week j and every combination (j,k) of weeks appears in the sample with the same frequency. This is proved in the appendix.

3. Models for response burden and attrition

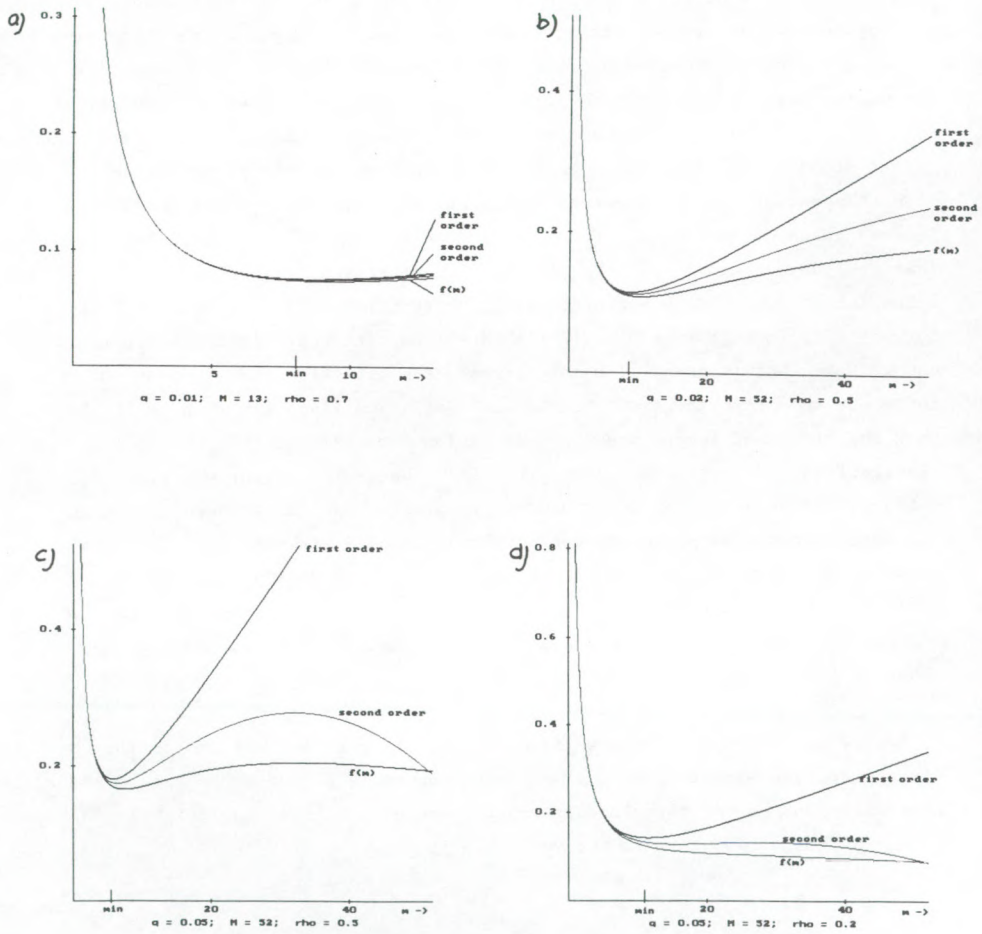
In this paper we will assume that there is a relation between attrition and response burden. We will study the behaviour of $\text{var}(X^{(t+1)}-X^{(t)})$ as a function of the attrition. The value of $\text{var}(X^{(t)})$ is trivial, because this variance is mimimal when we have as many as possible independent observations; consequently, it decreases when attrition increases. This makes the behaviour of $\text{var}(X^{(t)})$ less interesting not only from the substantive but also from the

statistical point of view. When we measure differences between intervals, however, it is well known that dependent observations may give higher precision than independent observations.

The response burden is defined by m, the number of measurements in a given period of M possible measurements. In model 1, such a relationship is already implied by its definition as each measurement causes a fraction q=1-p to leave the panel. Realistic values for q range from 0.1% to 5%, depending on the subject matter and the time horizon. Such values of q are sufficiently small to justify a linear approximation of formulas (5) and (6) for reasonable values of m, as is illustrated in figure 1. Throughout this paper we use linear approximations in order to change complex relations into simpler ones. In some cases optimum values for m are intractable for the variance functions that we study, but not for their linear approximations. Figure 1a shows the situation of 13 weeks in a quarter with 1% attrition each week and ρ=0.7 for formula (6). In this region $var(X^{(t+1)}-X^{(t)})$ and its approximations are almost equal. Note that in model 1, M only serves as a cut-off value. It does not influence the proportions of the variances as a function of m. Figure 1b shows that the first and second order approximation for q=2% and ρ=0.5 begin to diverge from say, m=15. This divergence is stronger in the unrealistic situation of 5% attrition and measurements over 1 year as is shown in figure 1c. Note that the second order approximation takes a maximum, just like f(m) between m=20 and m=40, whereas the first order approximation linearly goes to infinity. When also ρ is rather small, the first order approximation takes a minimum where f(m) and its second order approximation keeps decreasing, as shown in figure 1d.

Before we continue to discuss figure 1, we first give the formulas for the first order approximation as derived in the appendix (we omit the second order approximations as the formulas are complex and give little new insight). The linear approximation formula for $var(X^{(t)})$ is

Figure 1. Var($X^{(t+1)}$-$X^{(t)}$) as the function f(m); first and second order approximations for q in model 1.



$$\text{var}(X^{(t)}) \approx \frac{N^2 M^2 \sigma^2}{nm}\left[1 + (m-1)\rho - (m^2-1)\rho q/3 \right] \qquad (9)$$

The second term of this formula represents the well known cluster effect due to the fact that repeated measurements take place on the same respondents (Kish, 1965). The third term decreases this cluster effect because of the attrition: an expected proportion of q panel members is replaced every week. For m>0 (9) is a decreasing function of m. So, clearly, it is optimal to have as many measurements as possible. This need not be true if a replacement of a panel member comes with a price. This will be discussed in section 4.

For the variance of $X^{(t+1)}-X^{(t)}$, in the appendix the following first order approximation is derived

$$\text{var}(X^{(t+1)}-X^{(t)}) \approx \frac{2N^2 M^2 \sigma^2}{nm}\left[1-\rho + (2m^2+1)\rho q/3\right] \qquad (10)$$

By differentiating (10) with respect to m, it is shown that in this approximation the variance takes a minimum for

$$m_0 = \sqrt{\frac{3}{2}\,\frac{1-\rho}{\rho q} + \frac{1}{2}} \qquad (11)$$

So the optimum value of m_0 decreases with both ρ and q. This makes sense. When q goes to zero it is desirable to have a large m because attrition is low and we can profit from the fact that we have correlated single source data which are well suited for measuring differences. When ρ goes to zero, we have no reason to be careful to keep single source data for measuring differences, so we may just as well measure as often as possible, regardless of the attrition. If, on the other hand, ρ goes to one, the optimum value of m becomes less than one, so m=1 is the best possible value. From this one observation we can with certainty predict the values of $X_{ij}^{(t)}$ at other moments.

So far it was assumed that there is no direct relation between q and m. However, especially with burdensome surveys like a budget survey the respondents may drop out not only because of a questionnaire in the past, but also because of the prospect of the trouble they face in the future. A reasonable assumption is that q may have the form of

$$q = \lambda m^{\alpha}$$

This implies that the expected attrition in M weeks (q small) is of order $\alpha+1$, i.e. linear when $\alpha=0$, quadratic when $\alpha=1$ etc. To explore the consequences of this assumption we substitute q in equation (10). This leads to

$$\mathrm{var}(X^{(t+1)}-X^{(t)}) \approx \frac{2N^2M^2\sigma^2}{nm}\left(1-\rho + \lambda(2m^{\alpha+2}+m^{\alpha})\rho/3\right) \qquad (12)$$

This equation has analytic minima for $\alpha=0$ (when $\lambda=q$), $\alpha=1$ and $\alpha=2$. For $\alpha=1$ the minimum is

$$m_0 = \left(\frac{3}{4}\,\frac{1-\rho}{\lambda\rho}\right)^{1/3} \qquad (13)$$

and for $\alpha=2$ the minimum m_0 satisfies

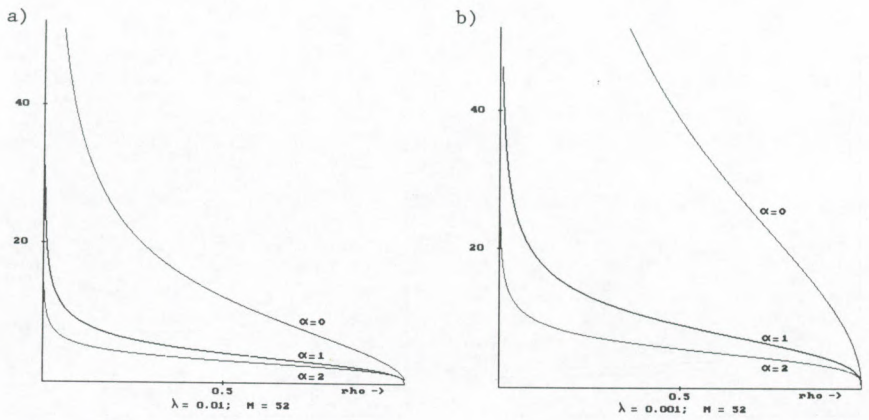
$$m_0 = \sqrt{-\frac{1}{12} + \sqrt{\frac{1}{144} + \frac{1-\rho}{2\lambda\rho}}} \qquad (14)$$

The interpretation of (13) and (14) is similar to (11), where λ has taken the role of q. By taking the respective roots, the values of m_0 decrease when α increases. In figure 2 the values of m_0 are plotted as a function of ρ for $\lambda=0.01$ and $\lambda=0.001$. For a small value of ρ the optimum value m_0 is high; for $\rho\to1$ m_0 goes to zero. Since the effect of response burden increases with α, m_0 decreases with α.

In the formulation of model 2 there is no direct causal relation assumed between q and m. This is reflected in the approximation formulas for the variances which are derived in the appendix.

$$\text{var}(X^{(t)}) = \frac{N^2M^2\sigma^2}{nm}\left(1 + (m-1)\rho(1-\frac{M+1}{3}q)\right) \qquad (15)$$

Figure 2. Optimum values of m as a function of ρ for α=0, 1 and 2.



and

$$\text{var}(X^{(t+1)}-X^{(t)}) \approx 2\frac{N^2M^2\sigma^2}{nm}\left(1 - \frac{m-1}{M-1}\rho(1-\frac{2M^2+1}{3}q)\right) \qquad (16)$$

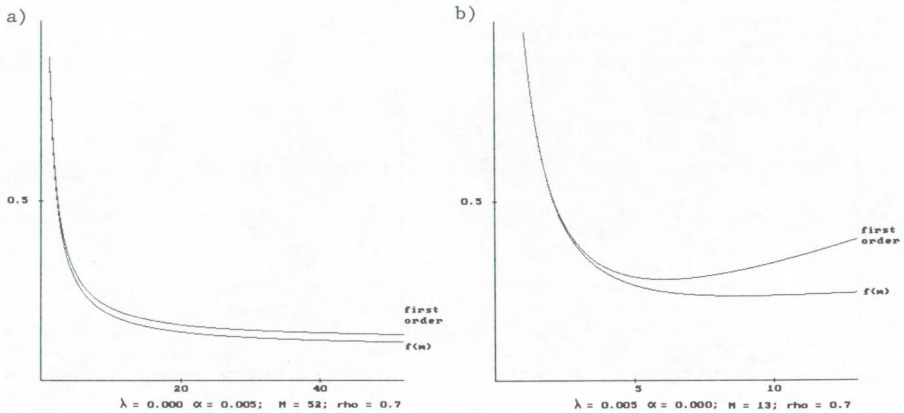
Both (15) and (16) are decreasing functions in m: we measure more accurately when more weeks are observed. This changes, however, when we assume a relationship between q and m. A reasonable assumption (and also one of the few tractable assumptions) is that q is of the form

$$q = \alpha + \lambda m.$$

This corresponds to the situation where other projects on the panel, which are considered to be given here, cause an attrition level α. This attrition level is raised by the budget survey with a term λm. Substitution in (16) and differentiation with respect to m leads to the optimum number of observed weeks

$$m_0 = \sqrt{\frac{3(M-1) + \rho\{3-\alpha(2M^2+1)\}}{\lambda\rho(2M^2+1)}} \qquad (17)$$

Figure 3. $Var(X^{(t+1)}-X^{(t)})$ as the function f(m); first order approximations for q in model 2.



It is obvious that m_0 decreases with λ: the more attrition associated with the budget survey, the smaller the optimum number of observed weeks. Given λ, m_0 also decreases with α.

Examples of formulas (8) and (16) are given in figure 3a for ρ=0.7. With α=0.005 and λ=0 the attrition is constant over time; it is clear that there is no minimum. The first order approximation (16) is close to (8) for m≤10. In figure 3b the results are given for α=0, λ=0.005 and M=13. Here both functions take a minimum. The first order approximation performs rather poorly for m>5. For this approximation we have $m_0 \approx 6$, whereas the real minimum is somewhere near 9. This minimum, however, is rather flat. Consequently, a suboptimal design does not give results that are much worse.

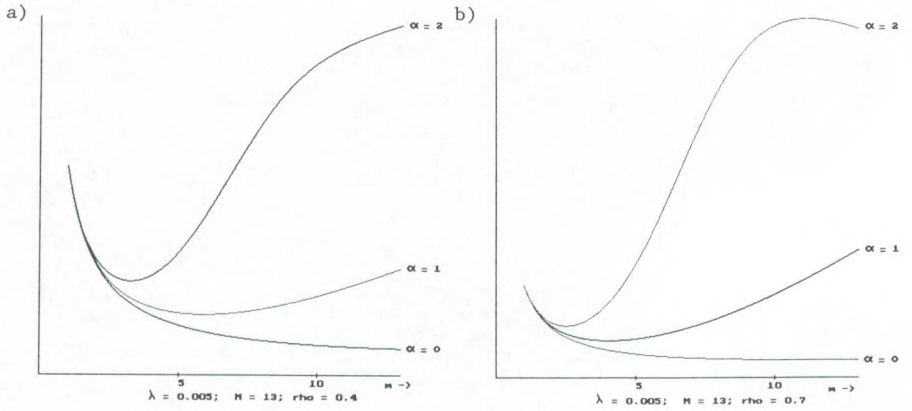
4. Practical considerations

In this section we abandon mathematical rigor, and we discuss what the practical implications are for decisions about the research design for the budget survey. We start with an issue that is traditionally considered in the classical textbooks on sampling theory (Cochran, 1977, Kish, 1965), namely cost considerations. For the calculation of optimum stratification or optimum cluster size usually a variance is minimized under some budget restriction involving costs per unit. For model 1 such a problem could look like: minimize (5) or (6), or a first order approximation like (10) under the budget restriction

$$c_1 n + c_1 nmq + c_2 nm = C \qquad (18)$$

where c_1 is the cost of finding and installing a new panel member, c_2 is the cost of processing one questionnaire and C is the total budget. Graphically, such problems can easily be solved, see figure 4. The trick is to write n as a

Figure 4. $Var(X^{(t+1)}-X^{(t)})$ with budget restriction (18) in model 1
$C=10^6$, $c_1=10^3$, $c_2=10$



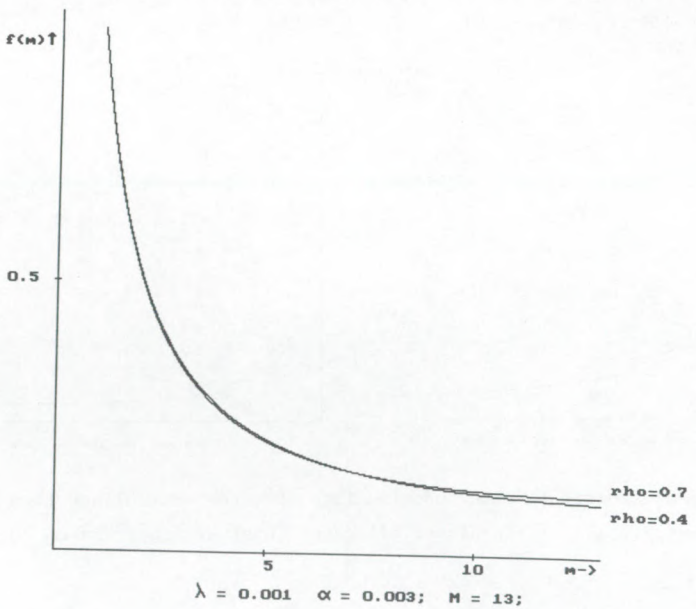
function of m, substitute this function into (5), (6) or (10) and then let the computer generate a plot. In almost all cases, however, there is no analytic

expression for the minimum m_0, especially when q also is a function of m. In many practical situations it is irrelevant to solve such a problem under budget restriction (18) because the sample size n is already fixed. This is especially true in case of a commercially exploited Telepanel, which is marketed as a panel of 1000 or 2000 households. This leaves us with the theory of the previous section.

In order to apply the results of section 3 many relevant data are missing. What *do* we know?

- the time periods j are weeks; the data are reported each quarter, so M=13.
- in the test period of one quarter, the respondents filled in the questionnaire twice (m=2)
- the attrition rate in the test period is approximately 0.5% each week
- other surveys may be a source of attrition; we do not know to what extent
- demographic developments are a source of attrition
- the correlations between measurements of the relevant product categories are between 0.4 and 0.7

Figure 5. $Var(X^{(t+1)}-X^{(t)})$ in model 2 as assumed in the Telepanel



So we guess that we have to use model 2 with α=0.003 and λ=0.001. In figure 5 we show $Var(X^{(t+1)}-X^{(t)})$ both for ρ=0.4 and ρ=0.7 as a function of m. In this case there is no optimum value m_0. Both functions are very similar. It is clear that in this situation the variance hardly depends on ρ. For m>7, the decrease in variance is very limited. Under the (heroic) assumptions given above, the most efficient decision therefore seems to be to have the respondents fill in the questionnaire for half of the weeks.

5. Summary and conclusions

In panel research, especially with a telepanel, research design is a difficult problem, because there are many factors to be considered. Even in a perfect world, where all respondents happily fill in their questionnaires without measurement error and without being bored by the many detailed questions about products, expenses, quantities etcetera, there are complex optimization problems because products are bought in different patterns, for which different sampling schemes are optimal. In this paper we restricted ourselves to global product categories for which it is reasonable to assume that they follow more or less the same patterns each week. We did not assume, however, a perfect world, but respondents who may become overloaded with their task. Under some very strict model assumptions we were able to calculate the variance of difference scores between two periods. Under even stricter model assumptions we could give expressions for m_0, the optimum number of weeks to be observed in a quarter or a year. In our practical situation it seemed reasonable to have the respondent fill in a questionnaire for one of every two weeks. The data on which these conclusions are based, however, are not very adequate.

There is still a world to be discovered in this field. In the first place the assumption that all respondents drop out of the panel with equal probability is not very realistic. It is more likely that there is a group of faithful respondents who, if it were up to them, would stay in the panel for life, and another group who drops out very rapidly. In the second place there are products that are bought with longer intervals than one week. When the purchasing process of such products is fitted to some statistical model, a

pattern of correlations between different weeks may come out that would not necessarily lead to intractable results. In the third place other classes of estimators (like composite estimators) may improve on the results we have given here. In the fourth place, models to correct for measurement error (which have not been considered in this paper), may have their impact on sampling design and estimation. Finally, time series models from which empirical Bayes estimators can be derived, can be used for optimum estimation of consumption.

References

Cochran, W.G., (1977), Sampling Techniques, 3rd edition, New York, Wiley.

Harrison, R., 1991, Respondent burden and respondent fatigue in the 1988-89 Australian Household Expenditure Survey, Paper prepared for the Workshop on Diary Surveys, February 1991, Stockholm, Sweden.

Hausman, J.A. and D.A. Wise, 1979, Attrition Bias in Experimental and Panel Data: the Gary Income Maintenance Experiment, Econometrica, 47, pp. 455-473.

Kish, L., 1965, Survey Sampling, New York, Wiley.

Lindström H.L., Lindkvist H. and H. Näsholm, 1989, Design and Quality of the Swedish Family Expenditure Survey, Proceedings of the Fifth Annual Research Conference, Arlington, U.S. Department of Commerce, Bureau of the Census, pp. 501-514.

Lyberg, L., 1991, Reducing Nonresponse Rates in Family Expenditure Surveys by Forming Ad Hoc Task Forces, Paper prepared for the Workshop on Diary Surveys, February 1991, Stockholm, Sweden.

Nevraumont, U., 1991, Evaluation of the Canadian Food Diary Survey, Paper prepared for the Workshop on Diary Surveys, February 1991, Stockholm, Sweden.

Ribe, M.G., 1991, A Study of Errors in Swedish Consumption Data, Paper prepared for the Workshop on Diary Surveys, February 1991, Stockholm, Sweden.

Saris, W.E., Prastacos, P. and M. M. Recober, 1992, CASIP: A Complete Automated System for Information Processing in Family Budget Research, in: New Technologies and Techniques for Statistics, Proceedings of the conference, Bonn, February 1992, pp. 80-87

Silberstein, A.R, and C.A. Jacobs, 1989, Symptoms of Repeated Interview Effects in the Consumer Expenditure Interview Survey, in: Kasprzyk, D., G. Duncan, G. Kalton, and M.P. Singh, Panel Surveys, eds. New York, Wiley, 1989, pp. 289-303.

Stasny, E.A., 1987, Some Markov-Chain Models for Nonresponse in Estimating Gross Labor Force Flows, Journal of Official Statistics, Vol. 3, No. 4, pp. 359-373.

Van de Pol, F., 1989, Reducing Panel Bias, A Review of Sampling Designs, Kwantitatieve Methoden 32, pp. 41-63.

Van de Pol, F., 1993, Weighting Panel Survey Data, Paper presented at Symposium on Analysis of Longitudinal Data, 1993, Tampere, Finland.

APPENDIX.

A.1 Design of model 2

Let the n individuals be randomly assigned to groups G_1, G_2, ..., G_L The group of individual i we denote by G(i). The groups correspond to the sets of weeks Γ_1, Γ_2, ..., Γ_L. Each set Γ_ℓ consists of m out of the M weeks. The design is such that for every week j we have $|\{i:j\in\Gamma_{G(i)}\}| = nm/M$ and for every combination (j,k) of weeks we have $|\{i:j,k\in\Gamma_{G(i)}\}| = nm(m-1)/(M(M-1))$ or, in each week with a fraction m/M of the respondents the budget survey is held and in each combination of weeks a fraction m(m-1)/(M(M-1)) of the survey is held. Now let $X_{ij}^{(t)}$ be the amount of purchases by household i in week j of period t if $j\in\Gamma_{G(i)}$. This definition is a slight generalization of the definition in the main text. Now let us calculate the variance and the covariance of $X_j^{(t)}$, defined according to (4).

$$\operatorname{var}(X_j^{(t)}) = \frac{N^2M^2}{n^2m^2} \sum_{i:j\in\Gamma_{G(i)}} \operatorname{var}(X_{ij}^{(t)}) = \frac{N^2M}{nm}\sigma^2$$

and for k>j

$$\operatorname{cov}(X_j^{(t)},X_k^{(t)}) = \frac{N^2M^2}{n^2m^2} \sum_{i:j,k\in\Gamma_{G(i)}} \operatorname{cov}(X_{ij}^{(t)},X_{ik}^{(t)})$$

$$= \frac{N^2M(m-1)}{nm(M-1)}\rho\sigma^2\rho^{k-j}$$

from which identities (7) and (8) are easily derived.

A.2 First order approximations of model 1

We start with model 1 from formula (5). Writing $q=1-p$ and $c_1=N^2M^2\sigma^2/n$ we can write for the variance of $X^{(t)}$

$$\begin{aligned}
\mathrm{var}(X^{(t)}) &= \frac{c_1}{m^2}\left[m + \frac{2\rho(m-1)(1-q)}{q} - \frac{2\rho(1-q)^2(1-p^{m-1})}{q^2}\right] \\
&= \frac{c_1}{m^2}\Big[\, m - 2\rho(m-1) - 2\rho\{1-(1-q)^{m-1}\} \\
&\quad + [2\rho(m-1) + 4\rho\{1-(1-q)^{m-1}\}]/q \\
&\quad - 2\rho\{1-(1-q)^{m-1}\}/q^2\Big] \\
&\approx \frac{c_1}{m^2}\Big[\, m - 2\rho(m-1) - 2\rho(m-1)q \\
&\quad + 2\rho(m-1)/q + 4\rho\{\tbinom{m-1}{1}q - \tbinom{m-1}{2}q^2\}/q \\
&\quad - 2\rho\{\tbinom{m-1}{1}q - \tbinom{m-1}{2}q^2 + \tbinom{m-1}{3}q^3\}/q^2\Big] \\
&= \frac{c_1}{m}\Big[\, 1 + (m-1)\rho - (m^2-1)\rho q/3 \qquad \Big]
\end{aligned}$$

In a similar way we can approximate the covariance of $X^{(t)}$ and $X^{(t+1)}$

$$\begin{aligned}
\mathrm{cov}(X^{(t)},X^{(t+1)}) &= \frac{c_1}{m^2}\,\rho(1-q)\{1-(1-q^m)^2\}/q^2 \\
&= \frac{c_1}{m^2}\left[-\rho\{(1-(1-q^m)^2\}/q + \rho\{(1-(1-q^m)^2\}/q^2\right] \\
&\approx \frac{c_1}{m^2}\left[-\rho\{\tbinom{m}{1}q\}^2/q + \rho\{\tbinom{m}{1}q - \tbinom{m}{2}q^2\}^2/q^2\right] \\
&\approx \frac{c_1}{m}(\rho m - \rho m^2 q)
\end{aligned}$$

This enables us to compute the first order approximation of the variance of $X^{(t+1)}-X^{(t)}$:

$$\mathrm{var}(X^{(t+1)}-X^{(t)}) \approx \frac{2c_1}{m}(1-\rho + (2m^2+1)\rho q/3)$$

A.3 First order approximations of model 2

We start with formula (7). Writing $c_2=N^2M\sigma^2/n$ we can write for the variance of $X^{(t)}$:

$$\begin{aligned}
\mathrm{var}(X^{(t)}) &= \frac{c_2}{m}\left(M + 2(m-1)\rho\{\frac{1-q}{q} - (1-q)^2\frac{1-(1-q)^{M-1}}{(M-1)q^2}\}\right)\\
&\approx \frac{c_2}{m}\left(M + 2(m-1)\rho[\frac{1}{q} - 1\right.\\
&\quad - \frac{1}{(M-1)q^2}\{\binom{M-1}{1}q-\binom{M-1}{2}q^2+\binom{M-1}{3}q^3\}\\
&\quad \left. + \frac{2}{(M-1)q}\{\binom{M-1}{1}q-\binom{M-1}{2}q^2\} - \frac{1}{M-1}\binom{M-1}{1}q]\right)\\
&= c_2\frac{M}{m}\left(1 + (m-1)\rho(1-\frac{M+1}{3}q)\right)
\end{aligned}$$

In a similar way we can approximate the covariance of $X^{(t)}$ and $X^{(t+1)}$

$$\begin{aligned}
\mathrm{cov}(X^{(t)},X^{(t+1)}) &= \frac{c_2}{m}\frac{m-1}{M-1}\rho(1-q)\{1-(1-q^M)^2\}/q^2\\
&\approx \frac{c_2}{m}\frac{m-1}{M-1}\rho\left(\frac{1}{q^2}\{\binom{M}{1}q-\binom{M}{2}q^2\}^2 - \frac{1}{q}\binom{M}{1}^2q^2\right)\\
&\approx c_2\frac{M}{m}(m-1)\rho\frac{M}{M-1}(1-Mq)
\end{aligned}$$

so for the variance of $X^{(t+1)}-X^{(t)}$ we find

$$\mathrm{var}(X^{(t+1)}-X^{(t)}) \approx 2c_2\frac{M}{m}[1 - \frac{m-1}{M-1}\rho\{1-\frac{2M^2+1}{3}q\}]$$

Ontvangen: 28-2-1994
Geaccepteerd: 7-6-1994