UNIVERSITY OF GÖTEBORG

Department of Statistics

RESEARCH REPORT 1982:2

ISSN - 0349-8034

HOUSEHOLD MARKET AND NON-MARKET ACTIVITIES - DESIGN ISSUES FOR A PILOT STUDY

by

TOMMY JOHNSSON

Statistiska institutionen
Göteborgs Universitet
Viktoriagatan 13
S 411 25 Göteborg
Sweden

HOUSEHOLD MARKET AND NON-MARKET ACTIVITIES - DESIGN ISSUES FOR A PILOT STUDY

by

TOMMY JOHNSSON

ABSTRACT

The design of a pilot survey of households is the subject of this paper. A few methods to collect data about consumption expenditures and time-use are to be compared. Among the design issues are the allocation of the sample on experimental groups, on strata and on days of the week.

1. Introduction

As part of the planning process for a major survey of the market and non-market activities of Swedish households, a pilot survey will be conducted.[1] This survey should fulfil several purposes. One is to accommodate a comparison between a few methods to collect data about consumption expenditures and time-use. Both diary methods and retrospective questions are considered. The pilot survey should also include tests of questionnaires and give an indication of the likely response rate in the main survey. It should also give estimates of the variances of various key variables to be used when designing the main survey. The first issue, i.e. the comparison of methods, has, however, dominated in the design of the pilot survey.

There are in particular two methods to be compared. The first is the diary-method, i.e. the respondent keeps a diary of time-use in various activities or the consumption expenditures during a specified period. The filled in diary forms are either collected by interviewers or returned by mail. The second method is that of retrospective questions. In a personal interview or a telephone interview the respondent is asked to recall time-use or consumption expenditures for one or several days.

Our budget constraint makes it impossible to include more than 300 observations in the total sample. For the same reason it was also decided to limit the pilot survey to three of 24 provinces in Sweden.

The main survey will be a subsample from a panel study of household incomes administrated by the National Central Bureau of Statistics, Hushållens Inkomster (HINK). In this way it will become possible to merge the detailed income and wealth information from the HINK files with new survey data. To image this aspect of the main survey it was decided that the pilot study

[1] The research program for this project is included in Eliasson, G.and Klevmarken, A. (1981).

should also be a subsample from HINK, but drawn from a panel which is not to be used in the main survey. The HINK panels are obtained as stratified random samples from the population of Swedish adults. The sample is stratified by household type and household income.

Section 2 gives criteria for the comparison of methods and specifies the problem in statistical terms. Section 3 deals with allocation problems and section 4 consists mainly of calculations which show that an improvement of the precision is needed. Sections 5 and 6 give ways of obtaining such an improvement and section 7 gives the final design along with additional conclusions.

2. Estimation and criteria for comparison

Let μ be the true mean of the consumption of a certain commodity and assume that we have two estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ obtained by two different methods. Our problem is to determine by how much the two estimates have to differ to establish a significant difference between the two methods at, e.g. the 10% level. This problem can also be put in the following way: How many observations are needed to detect a relative difference, $E(\hat{\mu}_1 - \hat{\mu}_2)/\mu$, of e.g. 10% at a certain significance level?

The mean consumption per day within stratum h is estimated by

$$\hat{\mu}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \bar{Y}_{hi} \qquad (1)$$

where n_h is the number of observations from stratum h,

$$\bar{Y}_{hi} = \frac{1}{t} \sum_{j=1}^{t} Y_{hij}$$ (mean per day and individual) and

t is the number of days sampled.

The variance of this estimator depends on both the variance between days for each individual and on the variance between individuals. Assuming SRS without replacement within each stratum and considering the estimate for a whole period as the most interesting one, one gets

$$V(T\hat{\mu}_h) = \frac{1}{n_h^2} T^2 \sum_{i=1}^{n_h} \left(\frac{s_{hi}^2}{t} \cdot \frac{T-t}{T-1} + s_h^2\right) =$$

$$\frac{1}{n_h^2} T^2 \left(\frac{\sum_{i=1}^{n_h} s_{hi}^2}{t} \cdot \frac{T-t}{T-1} + n_h s_h^2\right) \qquad (2)$$

where $s_{hi}^2 = \frac{1}{T} \sum_{j=1}^{T} (Y_{hij} - \mu_{hi})^2$, $\mu_{hi} = \frac{1}{T} \sum_{j=1}^{T} Y_{hij}$,

$s_h^2 = \frac{1}{N_h} \sum_{j=1}^{N_h} (\mu_{hi} - \mu_h)^2$, N_h is the stratum size and

T is the total number of days in the sample period.

The population mean for the whole period, $T\mu$, is then estimated by

$$T\hat{\mu} = \frac{T}{N} \sum_{h=1}^{L} N_h \hat{\mu}_h \qquad (3)$$

where $N = \sum_{h=1}^{L} N_h$ and L is the number of strata.

It is not possible to make assumptions about each individual variance, s_{hi}^2, so let $s_{hi}^2 = ks_h^2$, for all i and h, i.e. let the individual between days variance be proportional to the between individual variance. Having the same k in all strata, the variance of (3) is obtained by

$$V(T\hat{\mu}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \, V(T\hat{\mu}_h) =$$

$$\frac{T^2}{N^2} \sum_{h=1}^{L} N_h^2 \frac{s_h^2}{n_h} \left(\frac{k}{t} \cdot \frac{T-t}{T-1} + 1\right) \qquad (4)$$

If it is assumed that the two methods are applied to two independent samples and that the two methods only differ with respect to the number of days sampled, t_1 and t_2 respectively, the variance of the difference, $\hat{\hat{\mu}}_1 - \hat{\mu}_2$, becomes

$$V(T\hat{\mu}_1 - T\hat{\mu}_2) = \frac{T^2}{N^2} \sum_{h=1}^{L} N_h S_h^2 \left[\frac{1}{n_{1h}} \left(\frac{k}{t_1} \cdot \frac{T-t_1}{T-1} + 1\right) + \right.$$

$$\left. + \frac{1}{n_{2h}} \left(\frac{k}{t_2} \frac{T-t_2}{T-1} + 1\right) \right] \qquad (5).$$

The panel which is to provide the sample is in itself a sample from 1979, where our subsample of individuals living in 3 provinces will be drawn in 1982. For each stratum we need to know the number of individuals in the 1979 population and in the panel which live in these three provinces in 1982. The panel could easily be matched with a file of current addresses for the Swedish population, but for the 1979 population this was not possible. The tape with the 1979 frame is no longer available. The stratum sizes therefore have to be estimated. This was done by extrapolating the rate of change in "current" stratum sizes 1979-1981 in the following way,

$$\hat{N}_h^{82} = N_h^{81} \sqrt{\frac{N_h^{81}}{N_h^{79}}} \qquad (6)$$

where N_h^i is the number of individuals in the three provinces in year i of those who belonged to the sampling frame for the same year i.

3. Allocation between methods and strata

When comparing two methods one has to take into account the number of days being sampled. If there is no dependence between days and the sample is SRS without replacement, the variance minimizing portion to be allocated to method 1 is given by

$$w_1 = \frac{(\frac{k}{t_1} \cdot \frac{T-t}{T-1} + 1) - \sqrt{(\frac{k}{t_1} \cdot \frac{T-t_2}{T-1} + 1)\ (\frac{k}{t_2} \cdot \frac{T-t_2}{T-1} + 1)}}{(\frac{k}{t_1} \cdot \frac{T-t_1}{T-1} - \frac{k}{t_2} \cdot \frac{T-t_2}{T-1})} \qquad (7).$$

Calculations show that k has to be rather high to make w_1 diverge much from 0.5 and hence, with almost no loss in efficiency, the sample could be divided into two equal parts, one for each method.

In order to maximize the information given from a sample of fixed size the sample should be allocated between strata in the best possible way. The allocation in the HINK panel is not necessarily the best for our purposes. Since the costs of observations are the same in all strata and none of T, t and k depend on h the allocation is given by the usual Neyman formula,

$$n_h = n \cdot \frac{N_h\ S_h}{\sum_{k=1}^{L} N_k\ S_k} \qquad (8).$$

The S_h's are all unknown and to obtain estimates assumptions are needed. Consider the coefficients of variation within strata. They are not likely to vary much between strata as long as our interest is limited to relatively broad commodity aggregates. The c.v. depends on commodity and it is also likely that it depends on various household characteristics, in particular income. In general one would expect that the coefficient variation would increase somewhat with income. Data from the 1978 consumers expenditure survey indicate, however, that it tends to be relatively close to 1 for broad commodity aggregates. For this reason and for simplicity it is assumed that

$$c.v. = S_{jh}/C_{jh} = 1,$$

where C_{jh} is the total annual consumption of commodity j in stratum h. An optional design will not be very sensitive - to minor deviations from this assumption.

If Y_h is the household's total net income one gets the consumption ratio $c_{jh} = C_{jh}/Y_h$. c_{jh} has been estimated for a few commodities and for each stratum by the National Central Bureau of Statistics using the Family Expenditure Survey, see table 5. S_{jh} is then estimated by

$$\hat{S}_{jh} = \hat{c}_{jh} \cdot Y_h \tag{9}$$

The optional number of observations, n_h, according to (8) may be impossible to get. This is because the HINK panel contains an insufficient number of households in a few strata. When this problem arises, an iterative procedure is suggested. Let N_h^H be the number of observations in stratum h of the HINK panel. If $n_h > N_h^H$, for any h, then $n_h = N_h^H$ and the remaining observations, $n-\sum_h n_h$ are allocated once again. This goes on until $n_h \leq N_h^H$ for all h. For the resulting allocation, see Table 7.

4. Calculations

A preliminary considered design was to divide the 300 observations into four parts, one for each of four different data collection methods. Each subsample should, according to (7), be of equal size, ca 75 observations. In paired comparisons of estimated totals, one is interested in whether there is any difference between methods and how big this difference has to be in order to be found. Estimating stratum sizes and variances as suggested above and allocating according to (8) the following results are obtained for different assumptions made about the c.v.

Table 1: Variance of an estimated difference 75+75 observations

c.v.	$V(T\hat{\mu}_1 - T\hat{\mu}_2)$	Relative difference that could be detected at the 10% significance level
1.0	4556	41 %
0.8	2916	33 %
0.6	1640	25 %
0.4	729	16 %

$V(T\hat{\mu}_1 - T\hat{\mu}_2)$ from (5) with $T = 14$, $t_1 = 7$, $t_2 = 2$ and $k = 1$. $T\mu = 269$ according to the data used.

As the c.v. are more likely to be 0.8 - 1.0 than 0.4 - 0.6, these results indicate that the sample size for each method, 75 observations, is too small. It is not possible to increase the total sample size but an improvement could be obtained by dividing the sample into just two groups of 150 observations each. Repeating the calculations of Table 1 with this new assumption gives Table 2.

Table 2: Variance of an estimated difference. 150+150 observations

c.v.	$V(T\hat{\mu}_1 - T\hat{\mu}_2)$	Relative difference that could be detected at the 10% significance level
1.0	2278	29 %
0.8	1458	23 %
0.6	820	17 %
0.4	364	12 %

5. Allocation in time

Although the results of Table 2 are improved compared to those of Table 1, it is not satisfactory when a relative difference as big as 20 % could remain undiscovered. One possible way of obtaining higher precision is to look at the allocation in time. In the pilot study one is not interested in estimating the consumption or time-use for a long period but rather to make the number of days, T, as small as possible. The reason is obvious; the shorter period the less is the variance due to variation between days. If it is assumed that the diary method is administered for a period of one week and if the number of interviewers is taken into consideration, T must be at least 14 days. Two alternative designs for the diary method are considered. One, A, is to divide the period into two subperiods of one week and allocate half of the sample to each week. Another alternative, b, is to randomly distribute the seven days of book-keeping over the whole period. Although unrealistic, for our calculations these seven days do not have to be connected. This alternative is the same as the 7-day alternative in the previous section. We are now interested in comparing the efficiency of the new alternative a with that of the old alternative b. A simplification is to consider only one stratum. Alternative a will then give

$$V_a(\hat{\mu}_n) = V(\hat{\mu}_h^{'} + \mu_h^{''}) \qquad (10)$$

where $\hat{\mu}_h^{'}$ is the estimate of the first subperiod and
$\hat{\mu}_h^{''}$ " " " " " second " .

If independence between the subperiods are assumed and $T = t = 7$, (10) reduces to

$$V_a\ (7\hat{\mu}_h) = \frac{7^2}{n_h/2}\ (S_h^{'2} + S_h^{''2}) \qquad (11)$$

where $S_h^{'2}$ is the variance between individuals within the first subperiod and
$S_h^{''2}$ is the variance between individuals within the second subperiod.

To be able to compare this to alternative b above, the variance for the whole period, S_h^2, could be written

$$S_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (\mu_{hi} - \mu_h)^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} \left[\frac{1}{2}(\mu_{ih}^{'} + \mu_{ih}^{''}) - \right.$$

$$\left. - \frac{1}{2}\ (\mu_{ih}^{'} + \mu_h^{''})\right]^2 =$$

$$\frac{1}{4}\left[S_h^{'2} + S_h^{''2} + 2\text{Cov}(\mu_{ih}^{'}, \mu_{ih}^{''})\right] \qquad (12).$$

If T = 14, t = 7 and $S_{hi}^2 = k \cdot S_h^2$ this gives

$$V_b(14\hat{\mu}_h) = \frac{14^2}{4n_h} (S_h'^2 + S_h''^2 + 2\mathrm{Cov}(\mu_{ih}', \mu_{ih}'')) (\frac{1}{13} k + 1) \quad (13).$$

Hence, the ratio between (11) and (13) depends on k and the correlation between the two subperiods. If $S_h'^2 = S_h''^2$ this ratio is

$$\frac{V_a(7\hat{\mu}_h)}{V_b(14\hat{\mu}_h)} = \frac{2}{(1+\rho)\ (\frac{1}{13} k + 1)} \quad (14)$$

where $\rho = \dfrac{\mathrm{Cov}(\mu_{ih}', \mu_{ih}'')}{S_{ih}' S_{ih}''}$.

This means that if e.g. k = 1, p has to be close to unity to get $V_a(\hat{\mu}_h) < V_b(\hat{\mu}_h)$. Alternative b will thus in general be preferable. In practice, however, the days sampled must be connected. The variance of an estimator based on a randomly allocated period of 7 days depends on the inter-day correlation. Without additional assumptions about this correlation it is difficult to compare this alternative to the previous two.

6. Variance reduction by repeated measurement

Another approach to the problem of reducting the variance is through repeated measurement. The variance of the difference $T(\hat{\mu}_1 - \hat{\mu}_2)$ is, in general,

$$V(T\hat{\mu}_1 - T\hat{\mu}_2) = V(T\hat{\mu}_1) + V(T\hat{\mu}_2) - 2\mathrm{Cov}(T\hat{\mu}_1, T\hat{\mu}_2) \quad (15).$$

The two estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ were previously assumed to be independent and hence the last term in (15) vanished. A design with a positive correlation between $\hat{\mu}_1$ and $\hat{\mu}_2$, might reduce the total variance. If every respondent is exposed to both methods the two estimates are likely to be correlated. In order not to influence each other the methods must be separated in time and the order between the methods must be controlled.

In addition to the possibility of a positive correlation this design has the advantage of an effectively larger sample. Since every respondent is observed twice the comparison of the two methods would be based on 300 observations for each method.

The idea is as follows. Randomly divide the sample into two groups of equal size. Expose the first group to the diary method during week one and to retrospective questions during week two. The retrospective questions are assumed to cover two randomly designated days. The other group is treated in the same way except for the order of the two methods. This design gives two estimates of the mean consumption for each week and the difference between the two methods is estimated as,

$$T\hat{\mu}_1 - T\hat{\mu}_2 = \frac{T}{2}(\hat{\mu}_1' + \hat{\mu}_1'') - \frac{T}{2}(\hat{\mu}_2' + \hat{\mu}_2'') \qquad (16)$$

where $\hat{\mu}_1'$ is the estimate for the first week by method 1,
$\hat{\mu}_2''$ - " - second -"- 2,
$\hat{\mu}_2'$ - " - first -"- 2, and
$\hat{\mu}_1''$ - " - second -"- 1.

Since it follows from the design that the two groups can be treated as independent samples $\hat{\mu}_1'$ and $\hat{\mu}_2''$ are uncorrelated with $\hat{\mu}_2'$ and $\hat{\mu}_1''$. Since method 1 covers a whole week the only source of variability for $\hat{\mu}_1'$ and $\hat{\mu}_1''$ is the between individual variance, while the variance for the estimators of method 2 also depends on the within individual variance, i.e. the variance between days. We thus obtain

$$\begin{aligned} \mathrm{Var}(T\hat{\mu}_1 - \hat{\mu}_2) &= \frac{T^2}{4}\left[\mathrm{Var}(\hat{\mu}_1') + \mathrm{Var}(\hat{\mu}_2'') - 2\mathrm{Cov}(\hat{\mu}_1'\hat{\mu}_2'')\right] \\ &+ \frac{T^2}{4}\left[\mathrm{Var}(\hat{\mu}_2') + \mathrm{Var}(\hat{\mu}_1'') - 2\mathrm{Cov}(\hat{\mu}_2'\hat{\mu}_1'')\right] . \qquad (17) \end{aligned}$$

Assume now that there is no interindividual covariance between the estimates of the two methods, and also that the two subsamples are formed stratum by stratum in such a way that the sample size in each stratum is $n_h/2$. From eq:s (4) and (11) it then follows that

$$\mathrm{Var}(T\hat{\mu}_1 - T\hat{\mu}_2) = \frac{14^2}{4} \sum_h \left(\frac{Nh}{N}\right)^2 \left[\frac{s_n'^2}{n_h/2} + \frac{s_h''^2}{n_h/2} \left(\frac{k}{2}\,\frac{7-2}{7-1} + 1\right) + \right.$$

$$\left. \frac{s_h'^2}{n_h/2} \left(\frac{k}{2}\,\frac{7-2}{7-1} + 1\right) + \frac{s_h''^2}{n_h/2} \right] =$$

$$\frac{14^2}{2} \left(\frac{5k}{12} + 2\right) \sum_h \left(\frac{Nh}{N}\right)^2 \left(\frac{s_h'^2}{n_h} + \frac{s_h''^2}{n_h}\right). \qquad (18)$$

Calculations based on eq. (18) give Table 3.

Table 3: Variance of an estimated difference. Repeated sampling

c.v.	$V(T\hat{\mu}_1 - T\hat{\mu}_2)$	Relative difference that could be detected at the 10% significance level
1.0	876	18 %
0.8	561	14 %
0.6	315	11 %
0.4	140	7 %

Given the broad income classes used to form strata it would seem reasonable that the correlation between $\hat{\mu}_1$ and $\hat{\mu}_2$ is positive, at least for aggregates of commodities. This would then improve the precision even more. However, we cannot be certain of a non-negative correlation, which is a risk we would have to take.

7. Conclusions

The results of the previous sections indicate that a design with repeated measurement is likely to be preferred to a design with separate subsamples for each method. This would in particular be the case if there is a positive interindividual correlation between the two methods.

These results were obtained using a number of assumptions, the realism of which one might discuss. In addition, the precision we could expect with such a small sample as 300 individuals is not overwhealming. One should, however, keep in mind that this comparison of methods is not the only purpose for a pilot study. There are also other reasons for it.

References

Cochran, W.G. (1977: Sampling Techniques, 3rd ed., John Wiley & Sons, Inc., New York.

Eliasson, G. and Klevmarken, A. (1981): Household market and nonmarket activities, Research Report No. 12, from The Industrial Institute for Economic and Social Research (IUI).

National Central Bureau of Statistics (1978): The Family Expenditure Survey, Liber Förlag/Allmänna Förlaget, Stockholm.

Table 4: Strata in the 1979 HINK panel

Stratum No.	Socio-economic group	Total income in thousands SEK	Stratum sizes estimated according to (6)
1	Pensioners	< 38	132,141
2	"	≥ 38	113,706
3	Farmers	< 40	11,520
4	"	≥ 40	1,969
5	Employers	< 45	18,183
6	"	≥ 45	11,609
7	Households with children	< 38	8,272
8	- " -	38 - 125	198,235
9	- " -	≥ 125	44,420
10	Households without children	< 38	8,403
11	- " -	38 - 125	146,168
12	- " -	≥ 125	34,496
13	Single persons with children		47,415
14	Single persons without children	< 75	313,945
15	- " -	≥ 75	15,240

Table 5: Consumption ratios

Stratum No.	Food	Clothing and shoes	Housing	Leisure and culture
1	0.281	0.081	0.296	0.103
2	0.157	0.056	0.160	0.067
3	0.502	0.132	0.341	0.098
4	0.189	0.078	0.154	0.069
5	0.439	0.147	0.559	0.175
6	0.167	0.072	0.162	0.093
7	0.554	0.213	0.504	0.232
8	0.182	0.080	0.219	0.109
9	0.110	0.058	0.150	0.082
10	0.312	0.077	0.231	0.104
11	0.141	0.058	0.144	0.073
12	0.076	0.044	0.119	0.044
13	0.228	0.125	0.290	0.152
14	0.142	0.062	0.198	0.104
15	0.071	0.037	0.133	0.056

These consumption ratios were computed by the National Central Bureau of Statistics from the Family Expenditure Survey 1978.

Table 6: Average disposable income

Stratum No.	Disposable income SEK
1	24,599
2	44,489
3	20,475
4	47,690
5	20,355
6	51,961
7	23,640
8	56,433
9	84,796
10	30,488
11	53,346
12	81,860
13	38,128
14	27,904
15	48,664

From the Family Expenditure Survey 1978.

Table 7: Sample allocation on strata according to eq. (8)

i) Unrestricted

	Consumption commodity			
Stratum No.	Food	Clothing and shoes	Housing	Leisure and culture
1	35	25	31	23
2	30	27	26	23
3	5	3	3	2
4	1	1	1	1
5	7	6	7	5
6	4	4	4	4
7	5	4	4	4
8	77	83	79	83
9	16	21	19	21
10	4	2	2	2
11	42	42	36	39
12	9	12	11	9
13	16	17	17	19
14	47	50	56	62
15	2	3	4	3

ii) Restricted by the number of households in the HINK panel

	Consumption commodity			
Stratum No.	Food	Clothing and shoes	Housing	Leasure and culture
1	22	22	22	22
2	24	24	24	24
3	6	4	4	2
4	1	1	1	1
5	8	6	8	6
6	5	5	4	5
7	6	5	4	4
8	68	68	68	68
9	20	23	22	25
10	4	3	3	3
11	44	44	44	44
12	11	14	13	11
13	20	20	21	23
14	58	58	58	58
15	3	3	4	4