On optimal sampling designs for price index surveys

Susie Jentoft, Tora Löfgren, Anne Vedø and Li-Chun Zhang
Statistics Norway

The sampling design for a survey is a key step in ensuring the quality of statistics. Typically in price index surveys, businesses are the sampling units while goods or services are the statistical units to which prices are associated, giving rise to an indirect sampling situation. This paper explores a phase approach to indirect sampling design for price index surveys. The Norwegian Consumer Price Index (CPI) survey data provides a practical example.

Introduction

An optimal sampling design aims commonly at one of two objectives. Firstly, a design can be considered optimal if it minimises the variance of a chosen target estimator given the fixed sample size. Secondly, a design can be optimal if it minimises the sample size, or a chosen sample-size dependent cost function, subjected to the constraints on the estimation uncertainty. More generally, good sampling designs aim to strike a sensible balance between the production costs for the survey organization, the response burden on the business community, and the accuracy of the statistical outputs.

Typically in price index surveys, businesses are the sampling units while goods or services are the statistical units to which prices are attached. Moreover, a price index is a different parameter than the population totals or means that are often the default target parameters in a sampling design. The indirect nature of sampling and the lack of a finite-population sampling variance of the price index are the two major theoretical challenges in planning price index surveys.

The Consumer Price Index (CPI) is one of the most important economic indicators. The Norwegian CPI is based on many different data sources (Johansen and Nygaard, 2012). For the sample survey data, the prices of around 500 representative goods (excluding food) are currently collected monthly from around 2000 businesses. This provides the basis of data in this study.

In this paper we explore a *phase approach* to both types of optimal indirect sampling designs for price index surveys using survey CPI data. The variance of the computed price index is evaluated using a model-based framework. The sampling design controls the expectation of the model-based variance over hypothetical repeated sampling from the finite business population, which is often referred to as the anticipated variance.

Due to time and space limit, the presentation will be focused on the optimal design that minimises the anticipated variance of the price index for the survey sub-universe of the Norwegian CPI, given fixed

sample size of business units. The phase approach that incorporates the other type of optimal design will be outlined, and some challenges to its implementation will be noted.

Price index and model-based variance

We calculate the model-based variance of the survey CPI index as follows. All the goods are divided into the so-called elementary groups, denoted by g = 1, …, G. The survey CPI ($\hat{P}$) is a weighted sum of the Jevons index ($\hat{P}_g$), one for each elementary group. These are given as

$$\hat{P} = \sum_g w_g \hat{P}_g$$

$$\hat{P}_g = \left(\prod_{j=1}^{n_g} y_{gj}\right)^{\frac{1}{n_g}} \Big/ \left(\prod_{j=1}^{n_g} x_{gj}\right)^{\frac{1}{n_g}}$$

Where w_g is the weight for elementary group *g* which stands for the proportion of total expenditure for that group; x_{gj} is the base period price of item *j* in *g*; y_{gj} is the price of item *j* in the statistical period of interest; n_g is the number of price observations for items in group *g*.

Zhang (2012) provides the model-based variance for the three commonly used elementary index: Carli, Dutot and Jevons. For the Jevons index, the model-based variances are given as

$$\widehat{Var}(\hat{P}) = \sum_g w_g^2 \, \widehat{Var}(\hat{P}_g)$$

$$\widehat{Var}(\hat{P}_g) = \frac{\hat{\sigma}_g^2}{n_g a_g}$$

i.e. conditional on n_g, where $\hat{\sigma}_g^2$ is the estimated variance of items in group *g* given by

$$\hat{\sigma}_g^2 = \sum_j (z_{gj} - \bar{z}_g)^2/(n_g - 1)$$

$$z_{gj} = \log(y_{gj}/x_{gj}) - \log \hat{P}_g$$

and a_g is the adjustment factor associated with a Jevons index:

$$a_g = 1/\hat{P}_g^2$$

Anticipated variance

The indirect sampling setting corresponds in our case to a two-way classification: the elementary group *g* and the strata *h* of the business population, where the same group of goods may be found in businesses from different strata, while a business unit can provide prices pertaining to several elementary groups.

Basically we would like to allocate the sample among the strata in a way that will yield the highest number of price observations in groups which have the highest variance. In practice we do not know in advance how many price observations we will collect in each group, so n_g is a random variable. For

planning of the sample we create a matrix (b_{hg}) based on historic data, where each element gives us the average number of price observations in each stratum-group:

$$b_{hg} = \frac{E(n_{hg})}{m_h}$$

and n_{hg} is the number of price observations from the sample in stratum *h* of group *g* and m_h is the number of sampled businesses in stratum *h*.

We assume that, under stratified simple random sampling with stratum sample size m_h, the expected number of price observations in stratum-group (hg) is given by $E(n_{gh}) = m_h b_{hg}$. Substituting the resulting $E(n_g)$ into the model-based variance above gives us then an approximate anticipated variance, based on which we are able to use the sampling design to control the estimation uncertainty. One can employ a more refined approximation to the anticipated variance, which takes into account the variance of n_{gh} and the higher-order terms of the Taylor expansion of the model-based variance in n_g. We do not go into the details here.

Algorithms of sample size allocation for minimising the anticipated variance

We have tested a range of algorithms for obtaining sample allocation that minimises the anticipated variance. Local algorithms (MMA, COBYLA and AUGLAG with MMA) were tested using the package *nloptr* in R with internal and box constraints. Additionally we tested some global algorithms (DIRECT, MLSL and STOGO). The results however varied greatly both among the algorithms and even more so when various starting points were tested. This seems to suggest the difficulty in finding a global minimum directly. We are currently still investigating the possibility of using readily available software packages in the hope of finding a viable solution directly.

Sample allocation using a fill-up algorithm

Meanwhile, we explored a greedy algorithm which we refer to as *fill-up.* This operates by adding one business unit at a time to the sample, for which we choose the stratum that results in the most decrease in the target variance. Repeated evaluation of the target variance for each additional sample unit allows us to keep track of the final allocation achieved. The algorithm provides us full control over the total sample size. It is also straightforward to accommodate restrictions of maximum and minimum stratum sample sizes, the latter of which can simply be set as the starting allocation.

We have tested this approach with CPI data (Table 1). In all cases, the total sample size was set to 2127 which is an average sample size for the year 2013. The stratification variable was the four-digit industry (NACE) code containing 38 strata. The grouping variable was a goods (COICOP) code and contained 516 groups. Five different constraints on stratum sample sizes were tested:

1) A minimum sample size of one business unit per strata
2) A minimum 50% of current stratum sample size as an average number for all months in 2013
3) A minimum 50% and a maximum 150% of max(10, current stratum sample size)
4) A minimum 50% of proportional-to-size (pps, turnover in 2013) stratum sample size
5) A minimum 50% and a maximum 150% of max(10, pps stratum sample size)

Strata industry code	Restriction 1	Restriction 2	Restriction 3	Restriction 4	Restriction 5	Currrent allocation	Proportional allocation
4520	126	96	88	96	166	59	135
4532	1	39	47	20	20	78	39
4540	12	14	16	10	10	28	7
4711	40	96	96	319	319	193	638
4719	164	82	105	102	60	70	40
4724	15	21	38	4	14	42	9
4729	1	9	9	4	10	18	7
4730	89	72	119	99	146	87	198
4741	1	26	76	5	15	51	10
4742	1	26	26	3	10	51	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Variance	**1.10e-06**	**1.24e-06**	**1.61e-06**	**1.29e-06**	**1.64e-06**	**2.21e-06**	**2.34e-06**

Table 1. Illustration of allocated stratum sample size by fill-up algorithm with varying restrictions (1-5) and fixed total sample size (2127) for 10 strata. Anticipated index variance is given in the last row.

We have tested two variations of the fill-up algorithm. The first is to add 5 units at a time, one in each of the 5 best choices of strata. The second is a stochastic approach where one stratum is selected from the top ten strata (those which provide the most variance reduction). The selection is based on a probability proportional to size of the improvement in the overall variance. Both these approaches gave similar results to that presented in Table 1 and are therefore omitted here.

In addition we have experimented with the reverse of fill-up, where we start with the population stratum size (as if in a census of business), and each time subtract one unit from the stratum where the resulting variance increase is the least. Again, this gave results similar to those in Table 1.

Sample allocation by swap algorithm

As a greedy algorithm the fill-up does not guarantee to find the global optimum, which does exist over the finite space of all possible stratum sample allocation. As an alternative, as well as a means for exploring the plausibility of the fill-up sample allocation, we consider a tabu algorithm.

The basic idea to swap business units between two strata: to move one unit from one stratum to the other and vice versa. The current sample allocation is updated by the move that leads to a smaller target variance. We refer to this as the *swap* algorithm. The swap algorithm becomes a tabu algorithm if the constraint is imposed such that a previously examined sample allocation is not to be revisited. In this way, the algorithm generates a sequence of sample allocations that decreases monotonely in the target variance. The algorithm is terminated if a chosen amount of swaps do not yield an accumulated variance reduction larger than a threshold value. It is not guaranteed to reach the global optimum in finite time.

Different starting points can be used to initiate the swaps. For example, we can start with the current sample allocation in the Norwegian CPI. Or we can start with the allocation achieved by the fill-up algorithm. We have tested the swap algorithm when 100 attempts to swap have been reached without finding an appreciable better solution. We have also tested higher number of swaps, but 100 appeared to be enough when the strata are chosen randomly. Moreover, the results from the swap algorithm were

similar to the fill-up allocation described above, indicating perhaps that the results presented in Table 1 are close to the global optimum under the respective restrictions. We are currently investigating potentially more efficient choices of the strata to be swapped.

Minimising sample size by down-size algorithm

It is possible to incorporate the other type of optimal design as an extension of results achieved so far. Essentially one only needs to include a set of additional constraints on the variances. These may contain the anticipated variances of several sub-indices. Starting from the sample allocation obtained above, we reduce the sample size, say, one at a time, choosing the stratum that yields the least variance increase of the overall price index, provided all the resulting variances satisfy the chosen constraints. Also the stratum sample size constraints can be adjusted adaptively to reflect the fact that the total sample size is being reduced all the time. We refer to this as the *down-size* algorithm.

Notice that the down-size algorithm can as well be applied with the current CPI sample allocation as the starting point. But it seems plausible to build on the previous attempt at minimising the target variance. The resulting approach consists then of two phases: first fill-up and swap, then down-size and swap. In this way, one hopes to arrive at a sampling design that strikes an aforementioned sensible balance between production cost, response burden and statistical accuracy.

Concluding remarks

Our experience so far suggests that it may be difficult to pin down unequivocally the definition of either type of optimal sampling design, and to find the corresponding global optimum directly. In response to this we have developed a phase approach to indirect sampling design for price index surveys, which aims at striking a sensible balance between the conflicting central objectives of sampling design.

The main practical objective of our ongoing investigation is to automate this phase approach, given the target variance and the constraints on estimation uncertainty and sample size. An important issue is to understand better the reason or how the sample allocation varies according to the different constraints. A related, albeit even more difficult, issue concerns the relationship between the global optimum and the potentially many local optimums that are nearly globally optimal.

References

1. **Johansen, Ingvild og Nygaard, Ragnhild.** *Various data collection methods in the Nowegian CPI.* Oslo : Statistics Norway, 2012.

2. **Zhang, Li-Chun.** A model-based approach to variance estimation for fixed weights and chained price indices. *Official Statistics Methodology and Applications in Honour of Daniel Thorburn.* 2010, 149-166.