Statistiska centralbyrån
Statistics Sweden

Variance analysis for PPI and SPPI

Final technical implementation report

Jan-Eric Wingren
Price Statistics Unit, Statistics Sweden
June 2009

Contents

1. Introduction

The aim of the work presented in this report is to estimate the sampling uncertainty, the variance, in all the different parts of the surveys as well as total dispersion for PPI and SPPI. The estimated uncertainty will then be used to achieve as optimal an allocation as possible of the sample and an improvement in quality reporting. In order to determine optimal sample sizes in accordance with some type of Neyman allocation, information is needed on the costs of data collection and variance in different strata, as well as the relative importance of the different strata. Without a more in-depth analysis of the cost function, until now costs have been approximately estimated as being the same in all strata.

In order to obtain a standard picture of how variances have changed over time in PPI, a time series was created with price data for all products over the period 1992-2007. However, there were problems in merging data from two different calculation systems. Not only were the variable names different, but errors/inconsistencies occurred in the historical price register. In certain cases, however, corrections have been made to the price register for the products' index figures using the method "link to show no change". These corrections have been implemented in the material used for index calculations, but in many cases the inconsistencies have not been remedied, in the majority of these cases the weights are low. After corrections to the original material, a number of measurements remain with extremely large and extremely low index figures. These extreme values will be analysed by other improvement projects for PPI.

The limitations of SPPI mean i.a. that no variances are estimated for the survey on rental of premises, and also that only one year is taken into account due to the difficulties of obtaining time series data.[1]

2. Terms

The different concepts used by Statistics Sweden when calculating index numbers for PPI is the December index/link which is the index figure for December based on the values from the preceding December. Also the unchained index is used which is surveyed month year y with base December year y-1 alternatively for the SPPI index figure with the 4th quarter of the preceding year as the base.

Products refer to the goods/transactions/services for which prices are recorded monthly or quarterly.

The division which is the basis for the strata in SPPI and PPI, for which variance is to be estimated, is the Swedish Standard Classification of Products by Activity 2002 (SPIN 2002).

[1] SPPI's studies are calculated today in Excel and each year has its own workbook.

3. Different possible variance estimators

In order to be able to at least make rough estimates of variance, some simplifications have been made. The calculations were done as if the samples were **pure** probability samples, although in reality they are not. In addition, no account has been taken of the fact that the samples (both for PPI and SPPI) are multi-stage samples where, depending on methods used, the sample is formed in two or three stages. Finally, coverage is not taken into account (finite population correction factor) for the period 1992-2005 (applies to PPI).

With the existing conditions as a starting point, the work has focused mainly on the possibility of estimating weighted variance within the "rough" variance in PPI. Attempts were also made to calculate unweighted two-stage variances, but the problem of a large number of companies with just one product in the strata (Producer price index, home sales 76%, Export price index 78% and Import price index 89%) exists.
This leads to a situation where variance within companies cannot be estimated, and the contribution to variance which comes from variations in a company is underestimated.

3.1 Variance estimator for independent random sample

If no account is taken of the fact that each product has a specific weighting, variance can be simply estimated by using the classic formula for variance within a stratum.

$$\hat{V}(I_i) = \frac{\sum_{j=1}^{n_i}\left(I_j - \bar{I}_i\right)^2}{(n_i - 1)n_i} \times (1 - f_i) \quad (1)$$

Where n_i is the number of products/sample units in stratum i.

3.2 Variance estimator for two stage samples

A more advanced method of estimating variance is to divide each product's contribution to the variance in two stages. Not only a component between companies, but also a component within each company. For estimating the variance for a two-stage sample, the following relationship is used: [2]

$$\hat{V}(I) = (1 - f_1) \times \frac{\sum_{i=1}^{n}\left(\bar{I}_{i.} - \bar{\bar{I}}\right)^2}{n(n-1)} + f_1 \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(1 - f_{i2}) \times \frac{\left(i_{ij} - \bar{I}_{i.}\right)^2}{m_i(m_i - 1)}}{n^2} \quad (2)$$

$$\bar{I}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} I_{ij} \text{ and } \bar{\bar{I}} = \frac{1}{n}\sum_{i=1}^{n} \bar{I}_i$$

Where *n* is the number of companies in the survey, *m* is the number of observations in the company, i is the company, j is the product and f_1 and also f_2 is the finite population correction factor in companies in the population.

[2] See e.g.[19] p. 278 formula 10.15.

A simplified version of (2) where the sampling fraction/coverage for the specific product in companies is set to zero.[3] results in:

$$\hat{V}(I) = (1 - f_1) \times \frac{\sum_{i=1}^{n} \left(\bar{I}_{i.} - \bar{\bar{I}}\right)^2}{n(n-1)} + f_1 \times \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{\left(i_{ij} - \bar{I}_{i.}\right)^2}{m_i(m_i - 1)}}{n^2} \quad (3)$$

3.3 Variance estimator for weighted variance

One method which takes account of how much each observation affects the point estimates is the weighted variance. Here account is taken of the importance of each observation with respect to how large a proportion the observation accounts for in the total weight of the stratum.

$$\hat{V}_1(I) = \frac{\sum_{i=1}^{n} \omega_i^2 \times \hat{V}_1(\bar{I}_i)}{\sum_{i=1}^{n} \omega_i^2} \text{ where } \hat{V}_1(\bar{I}_i) = \frac{\sum_{j=1}^{n_i} \dot{\omega}_{ij} \left(I_{ij} - \bar{I}_i\right)^2}{\sum_{j=1}^{n_i} \dot{\omega}_{ij} \times n_i} \times \frac{n_i}{n_i - 1} \times (1 - f_i)$$

(4) and $\omega_i = \sum_{j=1}^{n_i} \dot{\omega}_{ij}$

$$\bar{I}_i = \frac{\sum_{j=1}^{n_i} \dot{\omega}_{ij} I_{ij}}{\sum_{j=1}^{n_i} \dot{\omega}_{ij}}$$

$i = strata$

$j = price\ observation$

$f = sample\ \ fraction$

The difference between simple random sample and weighted variance is that each observation here receives a weight $\dot{\omega}_{ij}$ instead of $\frac{1}{n_i}$. This relationship applies both to estimates of the average value of the stratum $\bar{I}_i$ as well as for estimates of stratum variance $\hat{V}_1(\bar{I}_i)$.

The fact that the expression is extended by $\frac{n_i}{n_i - 1}$ is because correction for bias needs to be done. $\left(\frac{\sum_{j=1}^{n_i} \dot{\omega}_{ij}}{\omega_i} = \sum_{j=1}^{n_i} \frac{1}{n_j} \right)$

3.4 Finite population correction

Finite correction is used to correct overestimates of variance occurring for samples which are drawn without replacement from a finite population. Finite correction is (by definition) $\frac{N - n}{N}$ alternatively $1 - f$.

[3] The assumption that the sample fraction/(coverage) in the company is zero tends to overestimate the variance, but if few of all possible products in the companies surveyed have price measures taken, this simplification becomes reasonable.

Where f is a sample fraction and can be defined as $\frac{n}{N}$ or as $\frac{v_k}{\sum_{k=1}^{N_i} v_k}$ where v is the company's product/import/service value in the stratum and k is the sampled company and $\sum_{k=1}^{N_i} v_k$ is the total value of the stratum $\left(v_k \neq \sum_{k=1}^{n_k} \dot{\omega}_{ijk}\right)$.

The sample fractions which are applied in this paper are of the latter type, where the company's value is totalled in the strata and divided by the values of the stratum for total products/imports/services.

3.5 Margins of error

Estimates of the margins of error are consistently at a 95% confidence interval. The interpretation of the margins of errors is such that:
With a 95% confidence factor, it can be said that the true value for the index number lies within the interval
$I - 1{,}96 \times \sqrt{\hat{V}(I)} < I < I + 1{,}96 \times \sqrt{\hat{V}(I)}$.
A more formal interpretation is that 100 test samples drawn from a population cover about 95 of the sample's confidence interval, the true value of I.

4. Results PPI

4.1 Sampling fractions PPI

In order to provide sampling fractions for PPI, existing sample frameworks for 2006 and 2007 will be used i.e. production/import values will be based on industry, export and import statistics for years 2004 and 2005. Export/Import/Production values from frames have been matched against existing samples for PPI. Unfortunately, not all products in the sample match the values of products/imports in the frame, and this creates the need for a key between the existing sample and existing frames which indicates what/which company product is actually represented. The key developed for i.a. weighting calculation is however not complete and only covers around 100 products which have not been matched against the frame before drawing a sample.

The reason that the frame and sample do not fully correspond is due to a number of different factors. Below is a list of a number of different reasons as to why the discrepancy with respect to the frames occurs.

4.1.1 Different years for organisation registration numbers/incorrect organisation registration numbers

Since the sample frames are at least two years old, this means that PPI may have, or has, updated the organisation number into an actual organisation number during price collection.

4.1.2 PPI measures a closely related customs code.

In a number of specific cases, it has not been possible to establish a product in the customs code asked for/used, instead a closely related customs code in the same stratum has been chosen to represent the sample unit required.

4.1.3 Statistics Sweden measurements etc.

In many cases PPI uses material from other Statistics Sweden surveys and other public sources to reduce the burden of response on companies.

4.1.4 Total coverage

Total coverage for 2006 and 2007 lies between 35-55% of the total production/import value, which is an underestimate of the true coverage since a large part of the sample does not match any value in the frame.

4.2 Variance estimates for PC

The finite corrections used for the estimates are calculated from the material that was used for weighting in 2007. Finite correction has then been applied to the whole time period. For PCs as a whole total coverage is approximately 77% (KN 84713000 + 84715090). If PCs on the other hand are divided into portable computers (KN 84713000) and stationary computers (KN 84715090) the coverages are 68% and 86% respectively.

Table 1. Margins of error for PCs

Year	Margin of error total	Margin of error Stationary	Margin of error Portable
2004	6.13	5.88	6.36
2005	5.34	5.03	8.42
2006	4.15	5.80	4.52
2007	6.82	4.91	18.03

Table 1 reports the margin of error for PCs as a total and also disaggregated to portable and stationary computers. For the period 2004-2006, the margin of error was somewhat constant and if stationary computers for 2007 are compared, the margin of error decreases to below five index units. On the other hand something unexpected occurs in the estimates for portable computers where the margins of error show a substantial increase. The explanation for the margin of error increasing substantially for portable computers is that the hedonic model has not succeeded in explaining quality/specification changes that took place during 2007.

However, it should be pointed out that the total margin of error is calculated as if portable and desktop computers were one stratum, and not divided into two. This leads to a situation where the result is not necessarily the same for computers as a whole, which would have been the case if the stratum for stationary and portable computers had been combined with their respective weights. The reason that there is no distribution between stationary and

portable computers is that the samples within the different strata are so small and there is a risk that the estimates as a consequence are sensitive to errors.

Based on the reasoning above, the average variance for personal computers was estimated as $\hat{V}(I) \approx 8{,}45$.

4.3 Variance estimates for prefabricated wooden buildings (MT74)

The finite correction which was used for estimates is calculated based on the material used for weighting in 2006. Finite correction has then been applied to the whole time period. Coverage, calculated in accordance with turnover for 2005, was 73.7%.

The result of errors for assembling prefabricated wooden buildings is shown in Table 2 below. The margins of error are somewhat constant over time and if monthly data is analysed, no clear pattern is evident in the distribution of price changes.

Table 2. Average margins of error for MT74.

Year	Margin of error
2004	0.86
2005	0.79
2006	0.73
2007	0.92

The average variance calculated for MT74 was $\hat{V}(I) \approx 0{,}18$.

4.4 Variation for total index, Sections A-D

The average margins of error for the unchained index figures[4] at total index, sections A-D, during the period 1992-2007 were the following: Producer price index, home sales +/- 0.7, Export price index +/- 0.7 and Import price index +/- 0.7.

The corresponding margins of error for the December index (12 month figures for December) became +/- 0,8 for Producer price index, home sales, +/- 0.8 for Export price index and +/-0.9 for Import price index, somewhat higher than the average for all months.

If the result is then compared with estimates from method 2, the results are fairly similar. The average margin of error for unchained indices for the total index during the period 1992-2007 using method 2 for Producer price index, home sales was +/- 0.8, for Export price index +/- 0.9, and for Import price index +/- 1.0 index units on unchained index figures. The corresponding figure for the December index was +/- 1.1 for Producer price index, home sales, +/- 1.1 for Export price index and +/- 1.2 for Import price index.

Irrespective of choice of method, the margin of error is on a par with the educated guess of +/- 1 index unit from earlier studies.

[4] December year preceding year=100.

4.5 Stratum variation

There is wide variation between strata. There are strata where margins of error are close to zero, despite the fact that finite correction has not been applied and there are strata which have very wide dispersion, despite the fact that finite correction has been applied. For strata with extremely low variance, measuring errors and incorrectly allocated samples (variance exclusive finite corrections) can be suspected. The lack of variance may be due to the fact that only inliers and imputation are covered in the estimates of stratum variance, for small stratum samples, variance is more difficult to calculate. However, the variance does not need to be incorrect since there may be a high degree of homogeneity in the stratum.

In the strata with extremely high variation, problems concerning homogeneity can be suspected to exist in the stratum. Also problems with product mix, too small samples and errors/shortcomings in quality values are possible explanations.

4.5.1 How large is the variation within the strata?

To examine the problems of heterogeneous strata /errors in allocated strata, it should be mentioned that variance estimates for specific sample strata for the period 1992-2007 show major variations between strata.

For specific months, the margin of error for a stratum is at levels of +/- 30 index units even though finite correction has been applied.

Another problem may be the case where the margin of error is zero or close to zero over a longer time period. How should this be interpreted? Probably problems are reflected by inliers etc and also small samples in strata in these estimates in those cases where finite correction has not been done and/or was not possible. Also transfer prices/internal prices may play an important role in connection with the occurrence of low variances. But since the values of export/imports for production are based on the price picture, the price observations are correct but they can make it more difficult to estimate "correct" variances in the strata affected.

4.6 Has the change in method used for forming the sample had an impact on the margins of error?

In a brief review of how the estimates have been affected by the change of method, it appears that there is no improvement in accuracy, however the margins of error can be more accurately identified as the probability sample is used.[5].

[5] Up to the studies in 2007, slightly more than 40% of the sample was updated and rotated in accordance with [1], [2] and [7].

One question which occurs is why the margins of error have not decreased when the sample has been more "correctly" drawn? However, there is no clear-cut answer to this, but some possible causes are presented below

4.6.1 Has allocation of the new sample been done incorrectly?

This hypothesis is correct since there have been weaknesses in the material preventing optimisation when forming the sample/sample size in the strata in accordance with the method presented in [2]. Examples of shortcomings in the material are the following:

- **The variance calculations have been old**
  The variance estimates used are simple random sample estimates produced for the period 1998-2003, where the median value of the variance has then been chosen.

- **The stratum has not been and/or is not optimally created.**
  It has been difficult to create a homogenous stratum, which represents sufficiently large production/import values. In addition, another aspect is that the stratum should also be appropriate for all three markets. This has given rise to some strata becoming highly heterogeneous and they really need to be divided up and also that some strata are very small in certain markets.

- **Allocation of sample between strata has taken place incorrectly.**
  The SAS program which allocates sample sizes within and between strata is intended for drawing a completely new sample, not for a rotating sample. A consequence is that sample sizes between strata must be determined manually before drawing a new sample in specific strata, which has not taken place correctly.

4.6.2 Advantages of changing sampling method

Although the variation over time is not affected significantly, the advantages and disadvantages of changing method are being considered. More comprehensive evaluations, however, can not be carried out, with regard to accuracy, before the whole sample has been rotated and allocation problems have been corrected.
Some of the positive effects from the change of method are discussed below.

- **Interpretation of margins of error**
  For subjective samples, variance cannot be interpreted. In cases where a probability sample is applied, inferences capable of interpretation can be calculated for the point estimates.

- **Rotation of sample**
  Earlier the sample was not rotated in accordance with the guidelines drawn up by Statistics Sweden. The result is that small players often have an excessively high burden of response and some strata have had problems with products that are not representative of the stratum.

- **Annual reallocation**
  In connection with annual allocation in specific strata, the stratum will have more representative sample sizes than before. In addition the frames are processed more optimally, which leads to a situation where over- and under coverage that existed in earlier years is substantially reduced.

4.7 How should the sample be allocated?

Allocating sample sizes between the three markets also needs to be reviewed and investigated in order to minimise the margins of error and maximise coverage. Overrepresentation/underrepresentation on the domestic market, however, can be explained by the fact that material for forming samples for SPIN 01-05 has many shortcomings, and for this reason no review of these strata has yet taken place.

Table 3. Sample sizes etc 2007

Market	Total weight 2007 in SEK millions	Actual number of measurements 2007	Average weight per observation	Number of measures, proportional allocation, n=4000	Number of measurements, Neyman[6] allocation M1, n=4000
Domestic market	810012	1279	633.3	1353 (+ 6%)	1252 (-2%)
Export market	846659	1050	806.3	1414 (+ 35%)	1642 (+ 56%)
Import market	738493	1332	554.4	1233 (- 7%)	1106 (-17%)
Total	2395162	3661	654.2	4000	4000

The results in Table 3 indicate that sample size for export markets is too low and needs to be substantially increased. One explanation for the low export sample is that Sweden has just a few players with large volumes of exports with few products in strata which makes it difficult to achieve the required number of measurements. A more detailed analysis of what the distribution between the markets should look like needs to be carried out before final allocation can be determined.

4.8 What is needed to increase the accuracy of estimates?

To improve accuracy for the index numbers in PPI, a number of measures need to be taken. As mentioned earlier, sufficient resources are needed to create as optimal a stratum as possible for SPIN/SNI 2007, which was started in connection with drawing the sample for 2009. In connection with forming the sample, the sample size was increased from 4000 to 5000 units. Not only does the project need to work further on the mix problem, but also to map and correct products where transactions are not equivalent over time.

[6] The cost function is made constant and similar for all sample units.

A positive concrete step concerning inliers is that as of 2008, it will be possible to list the products which have unchanged prices for 24 months or more continuously in connection with regular validation.

Also training PPI staff and providing them with greater product knowledge to be able to evaluate the quality of products as optimally as possible is necessary.

In addition, work also needs to be continued on mapping transfer/internal prices to try to obtain a clear picture of how large a proportion of the sample consists of transfer prices, and to ensure that the problems that transfer/internal prices contribute, are handled in a correct way.

4.9 How should the results be used?

- **Sample allocation**
  Results can be used to allocate the sample more optimally.

- **Validation**
  Results can be used for validation purposes to identify problem strata i.e. strata which have large and small dispersion.

- **Publishing**
  In addition to this report, a section should be written on product descriptions.

5. Results SPPI

Since SPPI is a new product and coverage has been low in service industries, it has not earlier had a total index. Today coverage is better, and since 2008, a total index has been calculated. Some variance estimates for total SPPI have, however, not been done apart from the choice of a few industries where it was possible to estimate relevant dispersion measures. The methods chosen for estimating variances are in accordance with methods 2 and 3 and for comparative purposes the simple random sample estimator is also reported. The coming section reports the results for the industries investigated separately, and provides a brief description of the methods that it would be appropriate to introduce.

5.1 Sampling fractions SPPI

In most of the industries surveyed in SPPI, the companies drawn account for between 60 and 90 percent of total turnover in the industry, and this is then used as a sampling fraction. This largely contributes to the relatively low margins of errors.

5.2 Results in figures

Table 4. Margins of error in SPPI

Industry	Simple random sample	Two-stage estimator	Weighted estimator	Sampling fraction
Hotels	2.83	3.05		
Cargo handling	3.21	3.02		0.58
Storage and warehousing	2.80	3.29		0.65
Renting of automobiles	6.86	6.87	2.15	0.70
Hardware consultancy services	3.84	3.60		0.44
Computer consultants 72.21	2.66	3.41		0.45
Computer consultants 72.22	2.32	3.40		0.35
Data processing and database services 72.3	3.51	4.99		0.71
Data processing and database services 72.4	0.00	0.00		0.42
Computer Maintenance	2.64	3.05		0.91
Computer Operations, Others	4.22	4.16		0.64
Legal activities	2.08	2.17	1.93	0.35
Financial consultants	2.09	2.22	2.02	0.40
Architects	2.11	2.24		0.16
Technical consultants	1.83	1.95		0.28
Contracting staff	2.78	3.07		0.64
Investigation and security	2.17	2.73		0.88
Industrial laundry	0.18	0.15	0.08	0.81

5.2.1 Hotels

The margin of error in the hotel survey is ~ +- 3 index units for both simple random sample and the two-stage estimates. The reason that the estimators differ so little is that the industry is somewhat homogenous, the sample has good coverage and the dispersion between companies is less than the dispersion within companies.

5.2.2 Cargo handling

The margin of error for freight handling is ~ +- 3 index units for both simple random sample and two-stage estimation. The reason that the estimators differ so little is that the industry is somewhat homogenous, the sample has good coverage and the dispersion between companies and dispersion within companies is similar.

5.2.3 Storage and warehousing services

The margin of error for Storage and warehousing services varies by about 0.5 index units between simple random sample and the two-stage estimates. The reason that the estimators differ is probably that the industry is

somewhat homogenous, the sample has good coverage and since it has not been determined how large the sampling fraction is within companies, it is possible that the variance has been somewhat overestimated.

5.2.4 Car-rental

The result for car rental is misleading in terms of estimates using simple random sample and two-stage samples. The reason is that one outlier accounts for nearly all variation in the unweighted case. The company which has this service had a relatively small market share and the change did not have such a large impact on the index. If a comparison is made with weighted variance, where account is taken of each product's weight, it can be stated that uncertainty concerning the estimates are not particularly alarming.

5.2.5 Hardware consultancy

The margin of error varies by about 0.25 index units between simple random sample and two-stage estimates. The reason that the estimators differ so little is probably because the industry is somewhat homogenous, the sample has good coverage, and also because it has not been determined how large the sampling fraction is within the companies, it is possible that some overestimate of variance has been done for the multistage estimator.

5.2.6 Computer consultants

The dispersion for SPIN 72.21 and 72.22 is fairly similar. For both industries, there is a large measure of variation within companies which explains why simple random sample estimates are about 1 index unit lower than for the multistage estimator. This can be partially explained by the lack of a measure for the sampling fraction in companies.

5.2.7 Operations, computer processing SPIN 72.3

The dispersion for SPIN 72.3 shows the same trends as for 72.2 i.e. the multistage estimator has greater dissemination than simple random sample, also in SPPI measures, the dispersion is large for SPIN 72.3 which may be because the sample design needs to be reviewed

5.2.8 Operations, computer processing SPIN 72.4

Dispersion during the measuring period was 0 which is not realistic. The lack of variation for this SPIN is due to that the sample is small and that the price observations are imputations. A review of this industry is probably needed and the sample needs to be increased so that estimates of price changes are not misleading. Estimates for SPIN 72.3 may be more appropriate as estimators of variance in the strata when allocating samples.

5.2.9 Computer maintenance

For SPIN 72.5, the simple random sample estimate is a method which gives the lowest dispersion and the reason for the estimates differing so little is probably because the industry is fairly homogenous, the sample has good coverage and also that since it has not been determined how large the sampling fraction is within the companies, it is possible that some overestimate of variance has been made for the multistage estimator.

5.2.10 Computer operations, others

For SPIN 72.6 the estimate is similarly independent of whether simple random sample or a multistage estimate has been carried out. However, the dispersion is relatively great ~+/-4 index units which indicates that the sample size is at its very minimum.

5.2.11 Legal services

For legal services, it is possible to estimate a weighted estimate, which also gives the lowest margin of error ~+/- 1.9 index units, coverage however is low (35%).

5.2.12 Financial consultants

For financial services, it is possible to estimate a weighted estimate which also gives the lowest margin of error ~+/- 2 index units, coverage however is low (40%).

5.2.13 Architects

For SPIN 74.201 the simple random sample estimate is the method which gives the lowest dispersion and the reason for the estimators differing so little is probably because the industry is fairly homogenous and also that since it has not been determined how large the sampling fraction is within the companies, it is possible that some overestimate of variance has been made for the multistage estimator. However, coverage in this industry is at the minimum.

5.2.14 Technical consultants

For SPIN 74.202 the simple random sample estimate is a method which gives the lowest dispersion and the reason for the estimates differing so little is probably because the industry is fairly homogenous, and also that since it has not been determined how large the sampling fraction is within the companies, it is possible that some overestimate of the variance has been made for the multistage estimator. However, coverage in this industry is at the minimum.

5.2.15 Contracting staff

For SPIN 74.502 the simple random sample estimate is the method which gives the lowest dispersion and the reason for the estimates differing so little is probably because the industry is fairly homogenous, and also that since it has not been determined how large the sampling fraction is within the companies, it is possible that some overestimate of the variance has been made for the multistage estimator.

5.2.16 Investigation and security services

For SPIN 74.6 the simple random sample estimate is the method which gives the lowest dispersion and the reason for the estimators differing so little is probably because the industry is fairly homogenous, the sample has good coverage and also that since it has not been determined how large the sampling fraction is within the companies, it is possible that some overestimate of variance has been made for the multistage estimator.

5.2.17 Industrial laundry

SPIN 93.011 is the industry which has the lowest margin of error of all. For the weighted estimate, the error lies in a range of less than +/- 0.1 index units and coverage of 81%. This indicates that the industry is highly homogenous or that the services selected are highly homogenous and have stable prices over time.

5.3 Conclusions

Although the two stage estimator has a larger margin of error than simple random sample in certain cases, it is to be preferred since it better matches the actual sample process and should thus be a better help variable for sample drawing and allocation. Where there is a weighted estimator, this is to be preferred since account is taken here of how much each product contributes to the change in the target variable. Another conclusion is that the work of developing the basic material for variance estimates for SPIN 2007 needs to be done. In connection with this, a structure must also be created to more easily enable the automation of estimates for forming samples in the future. In addition, estimates are needed over a longer time period, and not just for one year quarter 4. In some, margins of error will probably vary between the years. Investigation will also be needed into what is an optimal level for the margin of error for a specific SPPI industry based on the cost of measuring data in relation to the values which the industry should regulate.

6. Recommendations

The recommendations of the project are the following:

- πps-sampling should continue to be used for both PPI and SPPI.
- Sufficient resources be allocated to enable the creation of optimal strata should be appointed.
- Variances with stratum allocations in accordance with SPIN 2007 should be estimated. The variances for PPI in accordance with new strata divisions were estimated before forming samples in 2008. New estimates will need to be made before sample drawing in 2009, which on a larger scale will correspond with the distribution of stratum in accordance with SPIN 2007.

- If the distribution of sample units between Producer price index, home sales/Export price index/Import price index is to be optimal, an investigation should be carried out into optimal sample sizes and how these should be allocated between the markets.

- Industries where extreme margins of error occur should be reviewed, and this applies to both PPI and SPPI.

- Variances for the Producer price index and the Price index for domestic supply should be estimated. From the user perspective, it is interesting to know what the dispersion looks like for the Price index for domestic supply since this index recurs most often in contracts.

7. Proposals for further work

1. Estimate variances in accordance with SPIN 2007

2. Estimate variance in accordance with Bengt Rosén's methods for estimating πps variances in [9] and [10].

3. Allocate the sample with reference to industries that are important for NA's deflators. In these cases variance is also calculated using the base year which NA uses. A formula for recalculating from December to the preceding year as the base period.

$$V(I) = V\left(\frac{I_{(12,y-1),(m,y)} \times I_{(12,y-2),12,y-1}}{\frac{1}{12}\sum_{m=1}^{12} I_{(12,y-2),(m,y-1)}} \times \sqrt{\omega_y \times \omega_{y-1}} \right) = \omega_y \times \omega_{y-1}$$

$$\times 12^2 \times V\left(\frac{I_{(12,y-1),(m,y)} \times I_{(12,y-2),12,y-1)}}{\sum_{m=1}^{12} I_{(12,y-2),(m,y-1)}} \right)$$

   One problem is that NA should have data from SPIN 2002 up to 2011.

4. Calculate variance for all annual figures. Since the problem of non-response/sample changes between the years will occur (products do not remain long enough) possible methods will need to be discussed and investigated for estimating variance.

5. Estimate margins of error for monthly/quarterly links i.e. how large is the variation with the preceding month/quarter as a base.

7.1 Further work on PPI

1. Analyse stratum variance further for delivery of material to other improvement projects.

2. Structuring databases so that it is clear what is dummy data, satellite calculations, stock prices, building index, no sales/imports, etc.

3. Deliver variances regularly for future sample draws

4. Build a validation function in the existing system to identify strata/products where problems occur.

5. Try to estimate variances for as many satellite calculations as possible. An appropriate method could be accepting that variance be calculated for the last two years and using the result for the whole period 1992-2007. Appropriate satellite calculations to start with are e.g. the material for the oil price index, import of coal. However, some of the material will need a large amount of processing.

7.2 Further work on SPPI

- Create some type of database to facilitate variance analysis/estimates.

- Implement variance estimates in sample formation in accordance with some type of Neyman allocation.

- Deliver variances regularly for future sample draws.

8. References

[1] Delrapport Aktualisering av PPI-urval; (2005-12-21)

[2] The sample project, An evaluation of pps sampling for the producer and import price index.
SCB Bakgrundsfakta 2005:03.

[3] Elementary Survey Sampling
Fourth Edition
Scheaffer, Mendenhall & Ott
PWS-KENT Publicing Company

[4] Model Assisted Survey Sampling
Särndal, Swensson & Wretman
Springer-Verlag (1992)

[5] Beräkning av varianser för PPI-indexen
Dalén (1995b)
PM

[6] Åtgärder för att höja PPIs tillförlitlighet
Dalén (1999)
PM

[7] PPI sample report 2006; 2006-12-20

[8] Mixproblem i PPI - Lägesrapport
Soukkan (2007)
PM

[9] A user´s guide to pareto πps sampling
Rosén (december 2000)
PM

[10] On sampling with probability proportional to size
Rosén (1997)
Jornal of statistical planning and inference 62 1997
s. 159-191

[11] Kvalitetsjustering av ICT-produkter
Metoder och tillämpningar i svenska Prisindex i Producent- och Importled
Deremar, Kullendorff (december 2006)
PM

[12] Estimation of variances in multistage sampling
Hans Jönrup (1974)
Statistisk tidskrift 1974:5
s.431-436

[13] Variance estimation for a ratio in the presence of imputed data
David Haziza (december 2007)
Survey Methodology, Vol 33;2
s159-166
Statistics Canada

[14] Statistiskt meddelande JO49SM0803
Prisindex och priser inom livsmedelsområdet. Års- och månadsstatistik 2008:1

[15] Sampling Theory
Des Raj
McGraw-Hill Inc. 1968

[16] Probability and statistical inference
Fourth Edition
Robert v Hogg & Elliot A Tanis
Macmillan Publishing Company 1993

[17] Estimation in the presence of Nonresponse and frame imperfections
Second Edition
Sixten Lundström & Carl-Erik Särndal
Statistics Sweden 2002

[18] Elements of survey sampling
Tore Dalenius, April 1985
Sarec IBSN 91 8682604-2

[19] Sampling Techniques
Third edition
William G. Cochran
John Wiley & Sons, Inc. 1977
IBSN 0-471-1640-x

[20] Dokumentation av TPI's urval
Kamala Krishnan 2007
Internt PM

[21] Urvalet i TPI
Stefan Svanberg 2003
Internt PM