PCA/MFFM
Anders Norberg

Ny information om urvalsosäkerhet i kvalitetsdeklarationen för KPI

För information

Enheten för prisstatistik avser att förbättra informationen till användarna om den statistiska kvaliteten i KPI:s viktigaste delserier. I den nya texten redovisas skattning av 95 procentiga konfidensintervall för månatlig prisförändring, inflationstakt och månatlig förändring av inflationstakt.

Beskrivning av statistiken

SCB dokumenterar och sprider information om statistikens kvalitet i s.k. "Beskrivning av statistiken" på SCB:s hemsida. Mallen för dessa dokument har två avdelningar:

A. Administrativa uppgifter
B. Kvalitetsdeklaration

I avdelning A finns bl.a. kortfattade uppgifter statistikens användning och undersökningarnas uppläggning.

Kvalitetsdeklarationen i del B följer SCB:s felmodell:

1. Innehåll (relevans)
2. Tillförlitlighet
3. Aktualitet
4. Jämförbarhet och samanvändbarhet
5. Tillgänglighet och förståelighet.

Tillförlitlighet har osäkerhetskällorna:

2.1 Urval
2.2 Ramtäckning
2.3 Mätning
2.4 Svarsbortfall
2.5 Bearbetning
2.6 Modellantaganden

SCB anlitar för närvarande de internationella experterna Paul Biemer och Dennis Trewin i ett program för att ta fram s.k. kvalitetsindikatorer som ett led i att bl.a. förbättra informationen om statistikens kvalitet till användarna. I deras modell riskbedöms varje felkälla för att avgöra felkällans potentiella inverkan på statistikens kvalitet. Varje felkälla

värderas med hjälp av fem generella kriterier och produkten betygssätts efter varje kriterium enligt en femgradig skala - från "svag" till "excellent". Sammantaget görs det mellan 30-40 enskilda bedömningar per statistikprodukt. Kriterierna är:

a. Kännedom om risker avseende datakvalitet
b. Kommunikation till statistikanvändare om dessa
c. Tillgång till expertis för att minska riskerna
d. Efterlevnad av standarder och best practices inom området
e. Insatser som finns på plats för att minska riskerna

De föreslår ytterligare två osäkerhetskällor inom tillförlitlighet; Specifikation och Revideringar. Med specifikation avses avvikelse mellan det som undersökningen ska mäta och det som faktiskt mäts, ej att förväxla med mätfel. Slutsatser baserade på den skattade målstorheten kan därför vara felaktiga. Under Revideringar anges förväntat avstånd /relativt avstånd mellan preliminär och slutlig statistik som ett mått på osäkerheten i den preliminära statistiken. Eventuell revideringspolicy bör kommenteras.

Ny deklaration av osäkerhetskällan Urval

Inspirerad av arbetet med kvalitetsindikatorerna har ett nytt innehåll i deklarationen av osäkerhetskällan Urval tagits fram för Beskrivning av statistiken. Texten finns fortfarande endast på engelska. Följande dokument har används som underlag för att skapa en samlad bedömning av osäkerheten i tre av KPIs statistikserier:

Dalén, J. and Ohlsson, E. (1995): *Variance Estimation in the Swedish Consumer Price Index*. Journal of Business and Economic Statistics, Vol. 13, No. 3, 347–356

Dalén, J. (2001): *Urvalsosäkerheter för olika tidshorisonter i KPI*. SCB, arbetspapper

Norberg, A. (2004). "*Comparison of Variance Estimators for the Consumer Price Index*" 8th Ottawa Group Meeting - Helsinki - 23-25 August 2004

Nilsson, H., Ribe, M. and Norberg, A. (2008) "*Variansberäkningar KPI*" Projektrapport, SCB, 2008-04-10.

Den nya texten omfattar fyra sidor i det följande.

2.2.1. Sampling

Three types of sampling in the CPI
The three types of sampling processes applied in the CPI are the following:

1) A sample of retail outlets (shops, hypermarkets, restaurants etc.) in the CPI is drawn annually in May with a so-called rotated sequential Poisson sampling with sample probabilities proportional to size (orderly PPS samples). Approximately 20% of the outlets are replaced annually and another 10% are replaced due to changes in the population, 70% remain in the sample for the following year. The sample is drawn within the frame for the coordinated sampling system for economic statistics, called SAMU, which is set from Statistics Sweden's Business Register The retail outlets in the local price collection in the CPI are divided into some 40 strata by type of industry according to the current standard for Swedish industrial classification (SNI 2007). Samples for many centrally collected prices, for example electricity, health care and entertainment, are renewed to a minor degree annually.

2) Prices for pre-packed products of food, detergents and other daily necessities are collected from about 40 retail outlets in three groups of store chains. Three different samples of 400 precisely specified representative products in approximately 80 product groups are utilised, one for each group of store chains. The samples are drawn from statistics from the chain's cash register data systems. The samples are chosen randomly using sample probabilities proportional to sales values. Samples of products for price measurement at pharmacies (not pharmaceuticals, which are measured in a different survey), tobacco shops and health food stores are also selected using the same method. Samples of representative products are updated annually. These samples are changed more slowly than outlets however.

Generic product specifications are established centrally for a large share of the remaining local price collection; thus, judgemental samples are applied here. Sources for these samples include information from the household budget surveys. The person collecting the data (the interviewer) is then instructed to choose the best selling product (by volume) in the selected retail outlets in terms of the specification.

3) The third sampling process is the selection of specific product varieties within the sampled outlets. For 30% of the new outlets all products varieties are of course new in the sample but varieties must also be replaced continuously when they are out of stock.

The samples of outlets and products in the central price collection are drawn partly by using PPS samples and partly by a variant of quota or cut-off principle. For certain products, total surveys (census) are applied, that is, all products within the specific area are included in the sample.

The CPI basket weights for product groups and industries are also updated annually. These weight are based on several sources of information, most of which are sample surveys. The household budget survey, HBS, is one important source. Due to large error margins, several years of HBS data are aggregated. The errors in weight imply errors in the CPI statistics, this contributions have not been estimated.

Complexity in the CPI affects sampling error
In statistics, sampling error or estimation error is the amount of inaccuracy in estimating some value caused by only measuring a portion of a population (i.e. a sample) rather than the whole population. This amount of inaccuracy is commonly referred to as sampling error and expressed as confidence intervals.

The complexity of the CPI statistics implies complex structure of the sampling error. Dalén & Ohlsson (1995) states that the independent sampling of outlets and products yield a two-dimensional, cross-classified sample. A design based variance formula is derived by exploiting the general theory for cross-classified sampling. This can be applied to the annual link from the base period (December year, *y-1*) to each month in the current year (year, *y*, and month, *m*).

The sampling errors due to sampling of outlets and products have a constant impact on one calendar year at a time. Norberg (2004) finds that the third sampling process of product varieties generally contributes most to the total sampling error. This sampling error is also found to be least correlated over time.

Some of the most important CPI-statistics involve several annual links, for example the inflation rate which is computed as the change from year, *y-1*, and month, *m*, to year, *y*, and month, *m*. The sampling error for the change statistics involve sampling errors for two samples (for two years).

Estimation of sampling error in the CPI
Dalén & Ohlsson (1995) proposes an analytic approach for estimation of variance in a cross-classified sample design of outlets and products. This can be applied to the annual link from base period (December year, *y-1*) to each month in current year (year, *y*, and month, *m*).

Dalén (2001) uses approximations and reasoning to motivate the best estimates of sampling errors for various reported measures of CPI changes that comprise more than one annual link.

Norberg (2004) studies the character of variation in price changes using analysis of variance models. Variance estimators in analytical forms are compared to estimators based on re-sampling procedures and models. All three methods result in estimates of roughly the same magnitude. Re-sampling procedures make it possible to estimate the variance for complex functions of index links such as the inflation rate and change of inflation rate without extra assumptions.

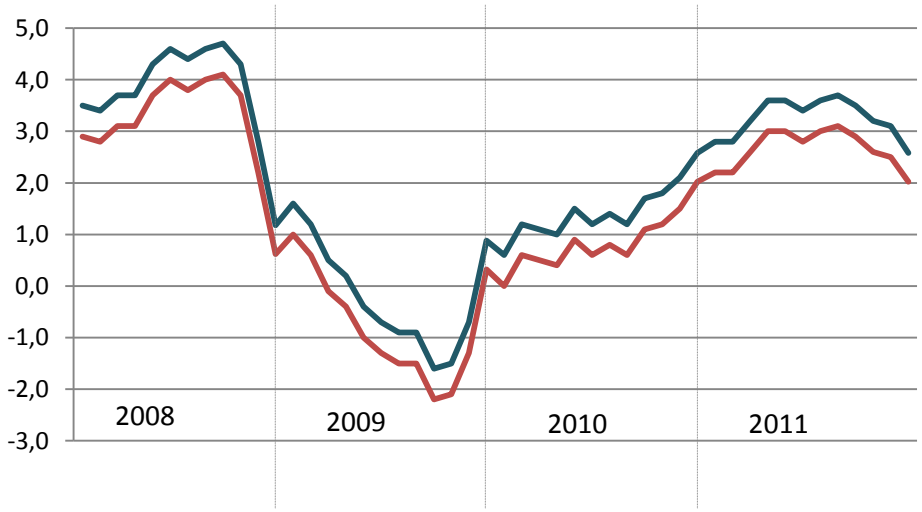
Nilsson, H. et. al. (2008) produces new estimates of sampling errors for the centrally collected product groups. These correspond to 46% of consumer expenditure.

Estimates of sampling error
Based on the four papers above, the sampling errors of the CPI-measures have been assessed and are given for the last years in the table below:

Table Estimated sampling errors, lengths of 95% confidence intervals 2012

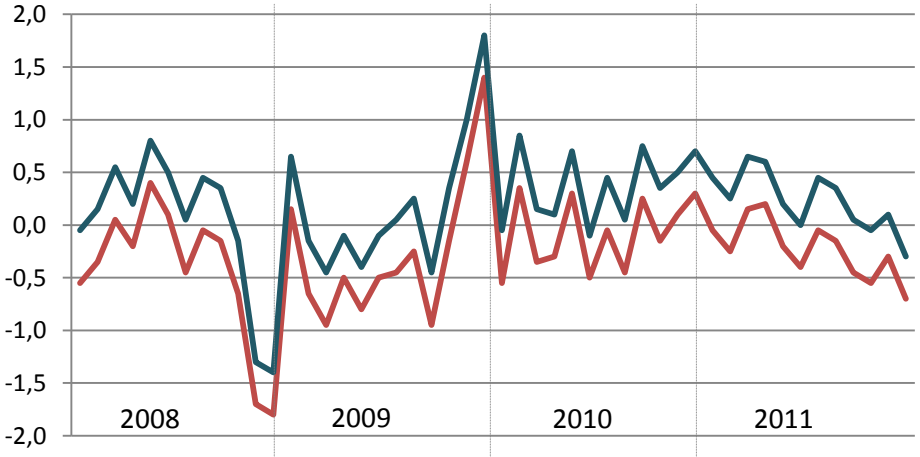
Statistics	Estimated length of 95% confidence interval	Comment
Monthly change	±0.15 - ±0.2	± 0.15 in April, May, June, November and ±0.2 other months
Annual change (inflation rate)	±0.3	Somewhat lower in December[1]
Monthly change of inflation rate	±0.2 - ±0.25	±0.2 in April, May, June, November, December and ±0.25 other months

Diagram 1 Inflation rate 2008-2011. 95% confidence interval



[1] The annual change from December to December is based on one and the same sample.

Diagram 2 Monthly change of inflation rate 2008-2011. 95% confidence interval



During 2011 the inflation rate change was "statistically significant" in March (up), April (up), October (down) and December (down). The inflation rate was +2.3±0.3% in December 2010 as well as in December 2011.

How sampling errors can be reduced

The general way to reduce sampling errors is to increase sample sizes. Statistics Sweden is making progress in using large amounts of cash register data (scanner data) received from retail chains. By using these data the sampling error for the area in question can be reduced as well as the costs for manual collection. For most other product groups more funding is needed to increase sample sizes.