ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Constructing variance estimates for the UK Consumer Price Indices: Progress update April 2024

Status: Work in progress

Expected publication: Alongside minutes

Purpose

1. The UK Statistics Authority, in approving the re-designation of CPIH as National Statistics, included a requirement to explore and publish estimates of quality (UK Statistics Authority 2016). Since then work to assess the sampling variance of the consumer price indices has been underway.
2. This paper summarises the progress made since April 2023.

Actions

3. Members of the Panel are invited to:
    - note the progress with variance estimation
    - note the methodological framework for estimating the variance of the CPI from the different components presented in Annex A. Any comments would be appreciated; but we are not expecting that the panel will review this methodology, which is more related to sampling than to index number theory.

Details

4. Since the previous update in April 2023 (APCP-T (23)02) we have:
    - reconsidered the question asked to the panel in April 2023 about how well the jackknife procedure as previously implemented captures the full complexity of the CPI price collection design, particularly the parts connected with locations, outlets and products. We now consider that a different approach is needed, explained in more detail in Annex A, to account for the multistage design. We are initially following an ultimate cluster approach, which makes important simplifications to the variance estimation, but for which the assumptions are at best met only tenuously. We will need to assess the impacts of these assumptions on the estimates in the future (which may require further, detailed calculations).
    - we have implemented the ultimate cluster approach approach (with some false starts); it is problematic because it requires recalculation of the full[1] CPI in each replicate, and therefore runs slowly.
    - set out a theoretical framework for estimating variances accounting for all the components of the design (see Annex A). This should now make it easier to understand

[1] That is, using all of the data available in the Secure Research Service, which is only prices from price collection, and therefore excludes some important components of CPI.

where the different calculations which have been presented so far (and those still in progress) fit into the overall estimates.

5. We continue to follow the plans for development of variance estimates.

Paul A. Smith & Francesco Pantalone, University of Southampton
April 2024

List of Annexes

Annex A	Variance estimation methodology for the UK family of Consumer Price Indices

Annex A

Variance estimation methodology for the UK family of Consumer Price Indices

Paul A. Smith
March-April 2024

1. Introduction

The Consumer Prices Index is a complex statistic which is composed of inputs derived from two principal sources, a bespoke price sample, and a weighting sample derived from the Living Costs and Food Survey (LCFS). In fact for the CPI family of indices the impact of the sampling in the LCFS is additionally obscured by a process of balancing the LCFS outputs through input-output balancing, with the LCFS proportions below this level reapplied to obtain the full detail used for weighting the CPI. For the legacy RPI calculations the LCFS data are used in raw fashion (but with some adjustments to which cases are included in the estimates), so the assessment of sampling variation is more straightforward.

2. Decomposition of the variance

At a high level we can use the law of total variance to decompose the variance of a general index estimator $\hat{I}$:

$$\begin{aligned}\operatorname{var}\left(\hat{I}\right) &= V_w\left(E_p\left(\hat{I}\,|\,\hat{\mathbf{w}}\right)\right) + E_w\left(V_p\left(\hat{I}\,|\,\hat{\mathbf{w}}\right)\right) \\ &= V_w\left(E_p\left(\hat{I}\,|\,\hat{\mathbf{w}}\right)\right) + E_w\left\{V_{III}\left(E_{I,II,IV}\left(\hat{I}\,|\,\hat{\mathbf{w}},\text{items}\right)\right) + E_{III}\left(V_{I,II,IV}\left(\hat{I}\,|\,\hat{\mathbf{w}},\text{items}\right)\right)\right\} \qquad (1a) \\ &= V_w\left(E_p\left(\hat{I}\,|\,\hat{\mathbf{w}}\right)\right) + E_w\left\{V_{III}\left(E_{I,II,IV}\left(\hat{I}\,|\,\hat{\mathbf{w}},\text{items}\right)\right)\right\} + \\ &\qquad E_w E_{III}\left\{V_I\left(E_{II,IV}\left(\hat{I}\,|\,\hat{\mathbf{w}},\text{items,outlets}\right)\right) + E_I\left(V_{II,IV}\left(\hat{I}\,|\,\hat{\mathbf{w}},\text{items,outlets}\right)\right)\right\} \qquad (1b)\end{aligned}$$

where subscript w denotes expectation/variance with respect to the design of the weighting survey (LCFS) and p denotes expectation/variance with respect to the design of the price collection. Within the price collection design, stage III refers to the selection of representative items, stage I to the selection of locations, and stages II+IV to the selection of outlets and products (prices), in line with the specification in O'Neill *et al*. (2017, box 8.1). It might, however, be better to regard the selection of representative items (stage III) as independent of the remaining stages; this kind of design has been called a cross-classified design (Dalén & Ohlsson 1995, Skinner 2015) and used as the basis for variance estimation in the Swedish CPI.

Broadly speaking, the first term in the last line of (1a) corresponds to the variance of the index due to the weights, the second term is the variance due to the selection of the representative items, the third is the variance due to the sampling of locations, outlets and prices. Note that we can continue this process further, and (1b) shows additionally a split into the components due to sampling locations and sampling outlets & prices respectively.

The expression for a Jevons-based index is not linear, but can be approximated by Taylor linearisation in order to obtain an expression which can be used with standard variance estimators. This process is relatively complex (and would need to be used in each of the variance components to be calculated in (1a) or (1b)). Instead we follow the strategy of most recent attempts and use a replication-based approach in order to avoid linearising the index estimator, which is made more complex because the statistic of primary interest is the inflation rate, the ratio of the index in a period to its value (usually) 12 months ago.

3. *Strategy for estimation of the different variance components*

We discuss the components of (1a) in order. We will need to make estimates of all these components, or to reason that they are negligible and can be ignored.

3.1. Variance component due to sampling for the weights survey

It has been known since the nineteenth century (Edgeworth 1888) that the weights have less impact on the quality of a price index than the prices themselves, but nevertheless they do have an impact, and this can be assessed by calculating the variance of the estimates of the weights derived from the LCFS. There are several ways in which such variance estimation can be achieved, but O'Donoghue (2017) used a bootstrap variance estimation approach to produce variance estimates for a year encompassing 2013Q3 to 2014Q2. We are able to use these as an indication of the variance, but there have been substantial reductions in the response rates in household surveys since this time, and it would be useful to be able to assess these differences. This would also allow us to assess the effects of the changes during the COVID pandemic on the variances of the CPI.

Proposal: Introduce a calculation for the sampling errors of the LCFS based on the data and design in the SRS.

O'Donoghue (2017) included some simplifications, in particular that the bootstrap replicates were not recalibrated to the known population totals. In principle this might be expected to increase the variability and to give an overestimate of the variances. But the falling response rates in household surveys since 2014 are likely to have had the opposite effect, to a considerably greater extent, so that we can expect that the current estimates of the variability of the CPI due to variance in the weights would be larger.

The independence of the design and collection of the LCFS from any of the price collection design is an advantage, as it means that we can calculate replicate sets of weights independently based on the LCFS without concern for any interaction with the price collection design. In addition, the LCFS has a true probability sample, so we do not need to make any assumptions about the sampling mechanism as long as we account for the features of the design. We need a set of replicate vectors of weights which can then be used to calculate the $V_w\left(E_p\left(\hat{I}\,|\,\hat{\mathbf{w}}\right)\right)$ term in (1a). [I think that we could just use the full price sample to obtain $E_p\left(\hat{I}\,|\,\hat{\mathbf{w}}\right)$, with no need for special calculation of the expectation as long as we believe the estimator is unbiased.] Note that in this process we are taking $\hat{I}$ as the statistic of interest, ie the 12-month change in the index.

3.2. Variance component due to sampling of representative items

The selection of representative items is a purposive part of the price collection design. A *de minimis* rule is used for which expenditures are to be included in the basket of goods for the CPI, giving us a cut-off sample if we consider that the CPI is to represent all of the items which are available to be bought by consumers. We will, however, ignore this component (which makes the design unmeasurable – ie that the variance does not exist); we can justify this either because the excluded parts are conceptually within the index but do not exhibit changes in the index different from those items which are included, or because the index conceptually excludes these items from its population.

Judgement is used to choose a suitable set of items which can represent the full range of items which is to be included in the index calculation. We can make different assumptions about this process:

a) we can assume that it is a fixed sample, which does not contain and random variation. In some senses this is close to the truth, since the choice of representative items is generally stable from period to period with only minor changes for basket updates one per year. In that case we could take $E_w V_{III}\left(E_{I,II,IV}\left(\hat{I}\,|\,\hat{\mathbf{w}},\text{items}\right)\right)=0$. However, because of the purposive nature of the sampling, we might expect there to be a bias relative to the true value of the index (in this bullet, meaning the index that would be obtained if all items were included, not just the representative ones). Ideally we should account for this bias in a mean squared error, but have no way to estimate it from the sample data under the assumption of this bullet.

b) we can assume that the choice of representative items is effective, so that whichever sample of representative items (among a set of potential representative item samples) was chosen, it would produce (approximately) the same estimates of price change. This argument is made by Skinner et al. (1996), as a result of which they regard the set of representative items as fixed (and therefore not contributing to the variance). This also results in $E_w V_{III}\left(E_{I,II,IV}\left(\hat{I}\,|\,\hat{\mathbf{w}},\text{items}\right)\right)=0$, but with a slightly different justification – and in particular that there is no systematic bias because the sampling is effective. I find this argument less convincing – if there are multiple potential samples, then we would expect them to shows some variation.

c) we can assume that the representative items are a pseudorandom sample from the population of available items – that is, that they originate from a probability sampling process even though we know that they don't. Then we can use any of the variance estimation methods; O'Donoghue (2017) used a jackknife variance estimation process. He additionally considered that representative items were divided (conceptually into strata) in two groups – those representative items that were in completely enumerated strata (usually with only one item in the population), so that they represented themselves, were certain to be included in the design, and therefore did not contribute to the variance; and those that were in strata with more than one potential item, which did contribute to the variance. We followed exactly this categorisation in calculating representative item variances in our paper Smith *et al*. (2023). We asked APCP-T for ideas for a principled approach to choosing which items would be regarded as self-representing, and they proposed identifying which prices had similar movements; no progress has yet been made with evaluating such an automated approach.

We assess that (c) is the best way of assessing the variance contribution in the CPI due to sampling of representative items. It acknowledges that some items are so important that they must be included in the index, so that no contribution to the variance comes from this source, but also reflects the differences in the index from the selection of representative items where there is some choice. The tendency to choose the most important items probably reduces the variance relative to a probability sample, but since these items would be selected with high probability, this is probably a fair reflection of the actual variability in the sample selection. We will assume that this variance component is appropriate for the sampling of representative items.

3.3. Variance component due to sampling of prices

This leaves the variance due to the price sampling, which comes in four stages – locations, outlets, products and days of the month. In principle we can continue the process of decomposing the variance conditional on the sampling outcomes at different stages, as in moving from (1a) to (1b), to give a series of expressions for the components of the variance due to each of the stages of this design. Then we could undertake a jackknife (or similar) replication process at each level to obtain an estimate of the necessary variance.

There are several ways in which these variances could be calculated. One procedure which is regularly used because of its relative simplicity is the ultimate cluster procedure (Hansen *et al*. 1953 vol. I pp.257-258 & vol. II pp.151-153). Under the assumption that sampling is with replacement, and that the sampling fraction is negligible then a procedure that works with the ultimate clusters (the most aggregated level of a clustered design) will capture the variance in the whole of a multistage design. Indeed, it will even capture the variance at lower stages which are due to non-random procedures (Wolter 2007 p48). We can apply this to jackknife replication by dropping each of these ultimate clusters in the jackknife approach (Wolter 2007, section 4.6). We discuss how to operationalise this procedure with the UK CPI price collection design in section 4. But first we need to consider how appropriate this approach is, and what alternatives might be available.

The jackknife is not appropriate for stratified designs, because it relies on the distribution of the jackknife replicates approximating the distribution of the statistic for which we need to calculate a variance. In a clustered design, different clusters are exchangeable, but in a stratified design the distributions within the different strata are different, so the elements in different strata are not exchangeable. Under certain conditions and where the appropriate weights for the different strata can be incorporated in the estimates, it is possible to develop a jackknife procedure which can cope with stratification through an amendment to the calculation of the pseudovalues (Wolter 2007, section 4.5). It also relies on the assumption that statistic of interest is a smooth function of the stratum means. This definition (not given, see Wolter 2007, p173) covers a wide range of statistics, and we assume that it also covers the price indices (and changes in price indices) in which we are interested.

The UK CPI price collection design is complex, with components with no stratification, and components with stratification by region, shop type, or region and shop type. In order to apply the ultimate cluster approach, we have to define different components of the design to be the ultimate clusters (and make the assumption that we have appropriately handled the stratification through the reweighting of the index when each ultimate cluster has been dropped). When the stratification is by region (1) or region and shop type (2), then the regions form the highest level of the design, and the ultimate clusters are the locations with the region strata. When stratification is by shop type only, or when there is no stratification, then the shop code forms the ultimate cluster. [This is the jackknife procedure that has been more recently (imperfectly) implemented in the SRS – accounting the different layers of the design. If we are prepared to make all the assumptions specified above, then this would account for all the stages of the complex price collection design – that is, it would generate the final term in (1a) and allow the full variance to be assessed.]

However, the assumptions are not really appropriate to this design; in particular the sampling fractions are not negligible, especially at the first stage where locations are selected from region strata. Some locations are selected with certainty. This is essentially the same as having two substrata, one completely enumerated and one sampled, within each region stratum. In the completely enumerated substratum there is no sampling (and therefore no exchangeability) of locations, and therefore we should fall back once more to shop type as the ultimate clusters within the jackknife procedure [but this is not currently/yet implemented]. Because the sampling fraction is relatively high in the sampled substratum and selection is actually *without* replacement, we should in principle incorporate a finite population correction. This would involve a different calculation (though it might be possible to include it in the calculation of pseudovalues). The selection of outlets and the selection of products also in principle need finite population corrections, but particularly for outlets the assumption of a negligible sampling fractions is more reasonable. There is little to no information on the number of available products for a particular representative item in a given outlet, but we might expect this also to be rather small and the fpc therefore to make an important contribution. In order to apply these fpc's we would need to estimate the different components of the variance separately, and apply the fpcs to each.

4. Implementation of the jackknife variance estimation

The ultimate cluster procedure described in section 3.3 is not so straightforwardly applied in the case of the calculation of the change in a price index. When an ultimate cluster is dropped in a jackknife replicate, it affects all of the price quotes in any of the outlets in that ultimate cluster (either a location or a shop code). In this replicate we then need to drop all of these price quotes, and then to recalculate the indices (and the inflation rate) using the remaining information (appropriately reweighted where necessary). There are about 24,000 ultimate clusters, and in each of these we need to calculate the CPI+CPIH from its component parts. This explains the relatively slow processing of the jackknife calculation using the ultimate cluster process. The latest version does not calculate each step in the CPI, but stores the complete-sample calculation, so that in any parts of the price collection which are unaffected by the jackknife deletion the index can be merely copied from the save version, saving some processing time.

5. Relation to previous experiments in the calculation of the CPI

The ultimate cluster approach is essentially the approach adopted by Sitter & Balshaw (1998) in their investigations using a simulated dataset in the late 1990s. It is similar to the approach adopted by Skinner et al. (1995) for investigations into the allocation of the price quotes in the CPI, though in their case they went further and calculated the variance components using the ultimate cluster approach to give the overall variance, and then subtracting the variances from the product sampling, calculated separately (and in both cases using Taylor linearisation rather than a replication based approach).

Jim O'Donoghue in unpublished work applied the jackknife right at the bottom level of the design by deleting single price quotes in each jackknife replicate. This is equivalent to assuming that price quotes are exchangeable within each lowest level index, which is much more reasonable, and calculates the variances assuming that the sample of price quotes in each elementary index is a simple random sample. It is effectively calculating $E_w E_{III} E_I E_{II}\left(V_{IV}\left(\hat{I}\,|\,\hat{\mathbf{w}},\text{items},\text{locations},\text{outlets}\right)\right)$ from a further-developed version of (1b), so one of the inputs to the overall variance calculation. Although this is a potentially very useful input, it does not capture the complexity of the higher levels of the design, so we need further investigations to make sure that these levels are properly accounted for in the variance calculations.

References

Dalén, J. & Ohlsson, E. (1995) Variance estimation in the Swedish consumer price index. *Journal of Business & Economic Statistics* **13** 347-356.

Edgeworth, F.Y. (1888) Memorandum by the secretary, Mr F. Y. Edgeworth, on the accuracy of the proposed calculation of index numbers. *Report of the British Association for the Advancement of Science* **1888** 188-219.

Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953) *Sample survey methods and theory*. Vol. I & II. Wiley, New York.

O'Donoghue, J. (2017) The effect of variance in the weights on the CPI and RPI. *Survey Methodology Bulletin* **77** 1-27.

Sitter, R. & Balshaw, R. (1998) Evaluation of bias and variance estimation of the RPI. Unpublished report to CSO.

Skinner, C.J., Holmes, D.J., Purdon, S., Haworth, M.F. & Beavan, R. (1996) Allocation of price data collection in the Retail Prices Index. Unpublished draft.

Skinner, C.J. (2015) Cross-classified sampling: some estimation theory. *Statistics & Probability Letters* **104** 163-168.

Smith, P.A. & Pantalone, F. (2023) Constructing variance estimates for the UK Consumer Price Indices: Progress update April 2023. *APCP-T(23)02*

Wolter, K. (2007) *Introduction to variance estimation*, 2nd edition. Springer-Verlag, New York.