THE IMPACT OF SAMPLE DESIGN ON THE PERFORMANCE OF THE SAMPLE GEOMETRIC MEAN AND RELATED ISSUES

1. Introduction

Inflation can be defined as "the process of a generalised and persistent increase in the level of prices". It can also be defined in terms of the general "fall in the value of money". As Sir Samuel Brittan has said "There is no one true and correct measure of inflation"[1]. Whatever the definition of inflation, it does not give a definitive steer on which prices should be included in an index designed to measure inflation except in the generality that you include prices that affect those people whose inflation experience you wish to measure. Indeed, even if one knew exactly which prices should be measured, there is still the issue of how these prices should be combined together into one number.

This paper concentrates on the main measure of inflation, the Consumer Price Index (CPI) or in the UK the Retail Prices Index (RPI). A Consumer Price Index can be regarded as a measure of the changes over time of the cost of a fixed "basket" of commodities, that is goods and services.

In the debate about potential bias in CPIs little attention has been given to the issue of bias resulting from sample estimation. This paper looks at the effect of sample design on the performance of the sample geometric mean, which is one method for combining consumer prices and the one which was advocated in the report by the Boskin Commission into potential bias in the US CPI. This paper does not come to a firm conclusion at this stage. The latter would require further investigative work based on historical price data. The main purpose of the paper is to raise the general issue of sample estimation.

2. Conceptual Issues

All price indices at the lowest level (i.e. for when there is no weighting data available) have to combine and compare price data for two periods. In the UK, indices are constructed by comparing the prices in the current month with the January base prices. These indices are known as **elementary aggregates**.

The general formula for an item index can be written as a weighted combination of elementary aggregates as shown below:

$$Item\ \ index = \sum_{r=1}^{R} \sum_{m=1}^{M} w_{irm} \hat{R}_{k01}^{irm}$$

where $\hat{R}_{k01}^{irm}$ is a measure of price change from time $t = 0$ to time $t = 1$ for item i within region r and shop type m using calculation method k. The three most commonly used calculation methods are a Ratio of Averages (RA), an Average of Ratios (AR) or a Geometric Mean (GM).

The three calculation methods are:

Ratio of Averages (k=1)

$$\hat{R}_{101}^{irm} = \frac{\sum_{lh} P_{1irmlh}}{\sum_{lh} P_{0irmlh}}$$

Average of Ratios (k=2)

$$\hat{R}_{201}^{irm} = \frac{1}{n_{irm}} \sum_{lh} \frac{P_{1irmlh}}{P_{0irmlh}}$$

Geometric Mean (k=3)

$$\hat{R}_{301}^{irm} = \left[\frac{\prod_{lh} P_{1irmlh}}{\prod_{lh} P_{0irmlh}} \right]^{1/n_{irm}}$$

where n_{irm} is the number of prices obtained for item i within location l in region r and shop type m and the sum or product over h is over those outlets where the item is on sale.

The population which is being estimated is different in each of these general cases.

Population Ratio of Averages (k=1)

$$\frac{\sum_{lh} P_{1irmlh}}{\sum_{lh} P_{0irmlh}}$$

Population Average of Ratios (k=2)

$$\frac{1}{N_{irm}} \sum_{lh} \frac{P_{1irmlh}}{P_{0irmlh}}$$

Population Geometric Mean (k=3)

$$\left[\frac{\prod_{lh} P_{1irmlh}}{\prod_{lh} P_{0irmlh}} \right]^{1/N_{irm}}$$

where i = item; l = location; r = region; h = outlet; $t = 0$, $t = 1$ for the two time periods and N_{ir} is the number of outlets in region r selling item i.

These are three entirely different ways of measuring price change. Thus comparisons between the estimators only make sense in terms of how well they estimate their respective population values. They are all attempting to measure average price change weighted by some measure of “market share”.

The sample Geometric Mean can be reformulated as below:

$$\hat{R}_{301}^{irm} = \left[\frac{\prod_{lh} P_{1irmlh}}{\prod_{lh} P_{0irmlh}} \right]^{1/n_{irm}} = \exp\left\{ \frac{1}{n_{irm}} \sum_{lh} \log\left(p_{1irmlh} / p_{0irmlh} \right) \right\}$$

Thus the GM can also be calculated by taking an average of log-ratios (natural log) and then applying the exponential. Similarly the population characteristic reduces to:

$$R_{301}^{irm} = \exp\left\{ \frac{\sum_{lh} N_{irmlh} \log\left(P_{1irmlh} / P_{0irmlh} \right)}{\sum_{lh} N_{irmlh}} \right\}$$

which shows its relationship to the estimator is similar to that of the average of ratios.

In general, ratio estimators, such as GM and AR, will have a bias of the order of 1/n. This becomes negligible as the sample size becomes large. In stratified sampling with many strata, the sample sizes may be small so the bias may become important.

3. Geometric Mean: Theoretical advantages and disadvantages

Irving Fisher[2] devised six tests which he argued that all reasonably constructed indices should pass (Appendix 1).

It is also important to consider the economic theory which defines a "true" price index as one which measures the cost of maintaining the consumer's standards of living based on the utility derived from purchases.

The geometric mean has been recommended, by some experts on index numbers, as a means of overcoming some of the problems associated with the use of AR and RA.

The following are generally considered to be the main advantages of using a geometric mean:

- the GM attempts to measure the costs of living and so reflects (albeit in a somewhat simplistic way) substitution due to relative price change.
- the GM balances the ratio of deviations of observations from it, rather than the sum of deviations as AR does. This means the GM gives equal weight to upward and downward movements within the elementary aggregate. For example, in an aggregate with five prices doubling and five prices halving, the GM will equal 100.
- price changes tend to be log-normally distributed and as such the GM is a good measure as it averages ratios and rates of change.
- the unweighted GM satisfies the first five of Fisher's test of an ideal index number.

The GM can be useful as it can be regarded as compensating for substitution bias due to shift of brands or varieties and outlets within item. The GM assumes that the expenditure share remains constant, so that if one price doubles while the others stay the same, the quantity purchased of the brand, variety or outlet represented by that price will halve. This assumes an elasticity of demand of 1 but actually demand is normally less elastic than this and so consumers will not necessarily reduce their purchases by half if the price doubles.

The following are some of the disadvantages of the GM:

- the GM is more difficult to explain, particularly to the "man in the street".
- the GM is a price weighted measure, with all price relatives being equally weighted as opposed to RA where more expensive varieties carry a greater weight.
- when there are unusually large price falls, the GM will be downward biased.
- the GM is optimal for some plausible models of consumer behaviour (and near optimal for others) but not for all behaviour. Thus in this respect it can be considered as broadly on a par with the RA as not having a totally satisfactory basis in economic theory.
- the sample GM is also a biased estimator of the population, as is RA, because they are both non-linear functions. In practice, the relative bias which results from the use of sample estimates will depend, amongst other things, on the design of the CPI sample. The sample design will also affect the relative impact on the index of using different formulae.

4. Geometric Mean: Impact of sample design

In the UK, initial estimates of the effect of changing the aggregation method from the current combination of AR and RA to GM for the HICP have been around -0.5%. This compares with France, Austria, Greece and Finland who all had estimates of around -0.1%. There may be a number of reasons for this large difference:

a) **Homogeneous versus Heterogeneous Items -** One of the main reasons that the French estimate is much smaller than that for the UK, may be the fact that the French define items to make them much more homogeneous than is the case in the UK. This means that in the UK a greater range of products, and therefore prices, are collected which means that, as the difference between the GM and the AR is proportional to the variance of price relatives, the UK will show a greater difference.

b) **Use of Stratification** - As mentioned previously, the difference between the GM and AR is proportional to the variance of the price relatives. Stratification in the sample design is normally used to reduce the variance within strata. In the UK, items are stratified by region and shop type (that is multiple versus independent shops), region only, shop type only or not stratified at all. The choice is dependent on the availability of reliable weights for the strata. The use of stratification will reduce the variance of the price relatives although as mentioned in the first point this will be confounded by the effect of how loosely items are defined. It will be difficult to disentangle the effect of the definition of items and the use of stratification.

c) **Estimator Bias** - The GM and RA are both non-linear functions which means they both have a bias of the order of one over the sample size. This is often ignored as the sample sizes are usually quite large so the bias will be small. This will not be the case if a very fine level of stratification is used so that sample sizes become quite small. Again, this will be confounded by the use of homogeneous versus heterogeneous item descriptions.

d) **Use of January Base Month** - In the UK, January is used as the base month which is the same month that there are large sales of clothing and consumer durables. The use of these exceptionally low prices for the base prices is likely to contribute to the difference between the GM and AR. In the case of AR, the use of a January base month can lead to an upward bias. This is because recovery from low prices will produce a few very large price relatives, which will lead to a large variance of the price relatives.

All four of these reasons are of interest in the context of the correct measurement of inflation and as far as the author is aware have not been addressed in the subsequent debate which arose from the Boskin Report. It is hoped to conduct further work into these issues during the course of investigations into the potential for bias in the UK RPI.

5. Geometric Mean: Estimation of Variance

A good sample design will minimise sample variance and produce an unbiased estimate of the true population. The exact form of the variance estimator will depend on the sample design. If there is a high level of stratification with high sample fractions within some strata, this will make the variance estimators much more complicated.

There are two general methods for variance calculation: direct calculations and replicated sampling methods. The Taylor Series linearisation is used to produce approximate formulae for estimating the variance of ratios in stratified sampling. This is necessary because ratios are non-linear functions, so the variance cannot be calculated directly. The calculations are often very complex and for this reason replicated sampling methods are often used instead.

Replicated Sampling Methods are commonly used for more complicated estimators. In these methods, a sub-sample is taken from the sampled data and $\hat{\theta}$ is calculated for that sub-sample. This is then repeated B times and the variance of all the $\hat{\theta}$ is then calculated. As long as B is relatively large and the sample design has been taken into account in the resampling then the variance of $\hat{\theta}$ should be an unbiased estimator of the variance of θ.

The advantage of these methods is that they can be used for any form of estimator and any sample design as long as the sample design is allowed for. A large number of sub-samples needs to be calculated to ensure high precision. The sub-samples need to be large enough to accommodate the sample design. The most common replicated sampling methods are the bootstrap and the jackknife. The jackknife involves deleting an observation, or with multi-stage sampling, deleting a primary sampling unit, and recalculating $\hat{\theta}$ with this value removed. A jackknife will only work on smooth functions of data, not for sample medians and quartiles. A bootstrap will work on non-smooth estimators but it is more computer intensive.

Dr Sitter, of Simon Fraser University, Vancouver, recently carried out work for the UK Office for National Statistics comparing the performance of the linearisation and jackknife variance estimators. For this work, a pseudo-population of data was set up with similar characteristics to the actual population of prices in the UK and of similar size. In order to capture the way in which prices between outlets, an ANOVA-type model was used. Actual data from six sections of the RPI were used to estimate the parameters in the model and then to build the finite pseudo-population. Four pseudo-populations were set up with slightly different parameters, referred to as cases 1-4 in the following tables. These results should be treated with some caution as they are based on simulated data. Four different aggregation methods were considered: AR, RA, GM and a combination of AR and RA as currently used in the UK RPI. The UK RPI currently uses AR for those goods and services where it is generally considered that the items are heterogeneous, for example, clothing and furniture. RA is currently used for those goods and services that have relatively homogeneous items, for example, bread and meat.

The simulated data was based on six sections of data: bread, footwear, audio-visual equipment, do-it-yourself (DIY), other household equipment and domestic services. It was assumed that the index was based solely on these sections. Only one of these sections currently uses RA for aggregation in the UK RPI, which explains why in the simulations AR and the combination of AR and RA behave similarly.

Table 1 shows the mean square error for the different aggregation methods using a completely self-weighting design. For these calculations, the estimated survey error of the RPI is

$$\text{MSE (RPI)} = \text{Var (RPI)} + \text{Bias (RPI)}^2.$$

Table 1: Relative Bias for Self-Weighting Sample Design

	Relative Bias			
	AR/RA	AR	RA	GM
Case 1	-8.38	-8.34	-5.68	-7.07
Case 2	-10.19	-10.09	-4.84	-8.01
Case 3	-9.43	-9.38	-6.82	-7.99
Case 4	-11.08	-10.92	-6.08	-8.80

The actual percentage change in the RPI varied between 1.09% and 1.36% for these cases, so a relative bias of 5% represents an absolute bias in the RPI estimate of between 0.05% and 0.07% while a relative bias of 10% represents an absolute bias in the RPI estimate of between 0.1% and 0.14%. In this exercise only local price collection is considered which comprises roughly half the weight of the index, if centrally collected prices were added, the bias would be halved as these have no sampling error.

Initial investigations suggest that there is little variation in the relative bias due to the different aggregation methods. In comparative terms, however, the percentage relative bias can increase dramatically depending on the sample design and where this occurs the choice of the different formula becomes more important.

To illustrate this, simulations were carried out to mimic a sample design where the largest locations were sampled with certainty and the number of small locations sampled was less than a self-weighting design would suggest. Each price collected was then assumed to be equally weighted with no weighting for the size of location. This is considered to be an approximation of the sampling scheme used in many countries where prices are collected mainly from the large locations.

Table 2 shows how the relative bias increases dramatically for this type of sample design.

Table 2: Estimated sample design with emphasis on large locations

	Relative Bias			
	AR/RA	AR	RA	GM
Case 1	-30.00	-29.73	-28.00	-31.42
Case 2	-34.08	-33.58	-30.82	-34.08
Case 3	-29.87	-29.64	-28.31	-29.87
Case 4	-33.85	-33.35	-31.31	-37.83

The purpose of this table is to show the importance of the sample design and the effect of not using a completely self-weighting sample design.

For these first calculations it was assumed that the selection of items was fixed, this means that only the effect of location and outlet sampling is included.

Table 3 shows a comparison of the variance estimators for a self-weighting sample design.

Table 3: Relative Bias of the Linearisation and Jackknife Variance Estimators

	Linearisation			Jackknife			
	AR	RA	GM	AR/RA	AR	RA	GM
Case 1	0.53	-15.33	1.22	0.53	0.53	9.91	0.31
Case 2	0.25	-17.72	1.27	0.14	0.25	11.75	0.09
Case 3	-0.63	-14.24	-0.17	-0.57	-0.63	7.23	-0.89
Case 4	-0.53	-17.04	0.26	-0.62	-0.53	9.66	-0.76

This shows that the Jackknife estimate performs extremely well, particularly for the RA. The reason for the large difference between the linearisation RA variance estimate and the jackknife is the impact of the bias. The previous results showed that the RA is typically very biased and this affects the linearisation variance estimate in particular.

The next stage was to consider the random selection of items. In the UK RPI the selection of items is purposive but can be approximated by a simple random sample without replacement within a sub-section. Within each sub-section items were then assumed to have equal weights (this is normally the case but not for every sub-section). The effect of randomly selecting items was simulated by creating a large set of possible items with the same mean and variance of price relatives as the real items. As different items were randomly selected, the weights were rescaled so that the weight within each stratum was always the same. Table 4 shows the relative bias of the different aggregation methods assuming a self-weighting sample design.

Table 4: Relative Bias Including Random Selection of Items

	Relative Bias			
	AR/RA	AR	RA	GM
Case 1	4.54	4.27	11.32	6.19
Case 2	6.80	6.34	18.47	11.88
Case 3	4.44	4.18	11.21	6.72
Case 4	6.91	6.49	18.38	12.86

Table 4 shows that the estimated RPI will be approximately unbiased for each of the aggregation methods.

Next, the variance of the estimator can be considered, this time including the variance due to item selection.

The variance is: $V(\hat{I}) = V_1 E_2(\hat{I}) + E_1 V_2(\hat{I})$

where V_2 refers to the variance under the random selection of outlets with items considered fixed and V_1 refers to variance under the random selection of items. The effect of having tight or loose item descriptions will affect the two terms in this equation and move the variance between the two terms whilst not affecting the overall variance.

Table 5: Relative Bias Using Linearisation and Jackknife Variance Estimators Including Random Selection of Items

	Linearisation			Jackknife			
	AR	RA	GM	AR/RA	AR	RA	GM
Case 1	11.31	8.42	11.91	11.21	11.31	11.57	11.73
Case 2	10.21	4.45	11.09	10.02	10.21	10.51	10.75
Case 3	10.96	8.06	11.60	10.83	10.96	11.30	11.42
Case 4	9.86	4.17	10.79	9.66	9.86	10.29	10.45

6. Further Work

The UK is currently embarking on a series of work to establish why the difference between the AR and the GM in the UK is so high. This work will involve investigating the effect of stratification on the GM. We will also look at the effect of item description and, therefore, price dispersion on the GM.

Further work is also being carried out on investigating the effect of the sample design on the index. Studies are being carried out on what is the best size measure to use when sampling towns. The ideal size measure would be total expenditure for each shopping centre but this is not available. Possible proxies for expenditure in a town are its population, the number of outlets or the number of employees. We hope to be able to make this analysis available at one of our future meetings.

Appendix 1

Irving Fisher's Tests for an Index Number

1. **Identity Test** - when one year is compared with itself the index shows "no change".
2. **Proportionality Test** - when all prices move in proportion, so does the index.
3. **Change of Units** - the index is invariant under any change in the money or physical units in which the prices of each individual product are measured.
4. **Time Reversal Test** - if the price of an item returns to the same level as the previous month, the index should also return to the same level.
5. **Circular Test** - the price index calculated backwards from t to 0 with the same weights as the forward index should be the inverse of the forward index.
6. **Factor Reversal Test** - the two index numbers between them account for the value change.

References

Allen, R.G.D., *Index Numbers in Theory and Practice*, (1975).

Brittan, Sir Samuel, *Minority Report on Treatment of Owner Occupiers' Housing Costs in the Retail Prices Index,* (1994).

Loynes, Prof. B., *Assessing the Representativity of the RPI*, (unpublished paper 1998).

Sitter, Dr R., *Evaluation of Bias and Variance Estimation in the RPI*, (unpublished paper 1998).

[1] Sir Samuel Brittan, *Minority Report on Treatment of Owner Occupiers' Housing Costs in the Retail Prices Index*, (1994)

[2] Fisher, Irving, *The Making of Index Numbers* (Boston, 1922)