On Sample Survey Designs for Consumer Price Indexes

Alan H. Dorfman
U.S. Bureau of Labor Statistics
2 Massachusetts Ave NE
Washington, D.C., USA
Dorfman_A@bls.gov

Sylvia G. Leaver
U.S. Bureau of Labor Statistics
2 Massachusetts Ave NE
Washington, D.C., USA
Leaver_S@bls.gov

Janice Lent
U.S. Bureau of Transportation Statistics
400 7th Street, SW
Washington, D.C., USA
janice.lent@bts.gov

Edward Wegman
Center for Computational Statistics
George Mason University
Fairfax, VA, USA
ewegman@galaxy.gmu.edu

1. Background

From start to finish, survey sampling for the sake of a Consumer Price Index (CPI) must rank among the most complicated of sampling enterprises. The population target is hard to pin down, the appropriate domain of items debated, the definitions of the raw ingredients – prices, quantities, items – ambiguous and subject to question. The ultimate estimator – the estimator of the all-items CPI – relies on data from at least two surveys, one giving prices, and one giving "weights." Below the level of "composite items" (or "item strata")– groups of items supposed homogeneous in their price movements – there is typically no way to properly attain the weights. Debate therefore goes on about the proper choice among various simple alternative estimators of the lower level price change, the "elementary aggregate indexes". The appropriate method of aggregating these price changes, using the weights, is subject also to debate.

There are two broad approaches to the sampling by which prices are collected: probability sampling and judgment sampling. The most commonly accepted approach to survey sampling in general requires injecting an element of randomness into the survey process and relying on this randomness to make inference about population characteristics of interest –probability or "design-based" sampling; see, e.g. Sarndal, et al (1992). This approach was not always taken for granted. Early in the 20th century "judgment" or "purposive" or "representative" sampling was considered a viable, and possibly better, option. More recently, the prediction-based school of Royall has challenged design-based assumptions; see e.g. Valliant et al (2000).

In the U.S., all CPI-related surveys are carried out using complex probability sampling techniques. Around the world, most CPI's are constructed from judgment samples, in which experts in the different item strata choose broader or narrower classes of items for which field representatives collect prices. The fundamental reason for this is the difficulty of getting all the data one needs on the plethora of items sold, and the places where they are sold, to make probability sampling feasible.

The interesting fact is that there has been very little assessment of the relative accuracy of the different approaches to sampling. Indeed it has not been clear that it is feasible to make such a comparison. The underlying population price index, for even the smallest of countries, involves so many transactions on so many items in so many places as to be inaccessible. Moreover, the population of items on the market is in a constant state of flux, complicating the application of traditional population index formulas. How then can one judge the relative closeness to "truth" of different sample-based estimates? Furthermore, not even sample information is available for a key ingredient of the population index--namely the *quantities* of item sold –so even artificially constructing a population for test purposes from sample data has not been feasible.

The relatively recent availability of *scanner* data, in the U.S. and elsewhere, presents an unprecedented opportunity for testing sampling approaches and estimators. These data include prices *and* quantities sold, typically on a weekly basis, of *all* the items sold in a given category, within a large sample of outlets having scanner devices. Such data may be used to construct realistic populations of transactions, for which the true price index is *known*. We can then use various methods to sample from this population, construct different index estimates of interest, and compare the results to the known population parameters. One such study, described by deHaan et al (1999), seems to show that "cutoff sampling," the sampling of the few largest (in terms of revenue generated) items in the population, outperforms two important design-based approaches: simple random sampling (*srs*), and probability proportional to size sampling (*pps*) (where the size measure is, again, revenue).

One difficulty in any such study is the task of maintaining a "level playing field." If one sampling method, for example, makes use of (population) information that might not actually be available in practice, while another does not, the comparison of methods is undermined. Similarly, if one method provides only one sample or a very few samples, and another provides thousands, special precautions are needed in comparing the two; indeed, such a comparison might require serious qualifications. Given the complexity of the sampling and estimation methods used in price index computation, it is not surprising that these and many other difficulties complicate experiments designed to compare various methods.

Note on the target indexes. There exist countless formulas for calculating price indexes between one period and another. Different indexes are compatible with different assumptions regarding the "average" consumer's buying behavior in response to price change. The "fixed market basket" indexes, the commonly employed Laspeyres, and the Paasche, are compatible with the assumption that consumers continue to purchase the same items in the same quantities regardless of changes in relative prices. The Laspeyres index projects the period 1 quantities forward to period 2, while the Paasche applies the period 2 quantities to period 1. The geometric mean (or "geomean") assumes that consumers adjust the quantities they purchase in such a way that the expenditure share for each item remains constant across time. The "superlative" Fisher and Törnqvist index formulas, which rely on quantity (or expenditure share) information for both periods, do not require these assumptions. The debate on the all items target index usually comes down to choosing between the Laspeyres and one of the superlative indexes.

2. The Present Study

The data source for the present study is a scanner data set for breakfast cereal purchased in the years 1995 through 2000 in three separate but contiguous sections of the New York metropolitan area. The data set was purchased from the Nielson Co. by the U.S. Bureau of Labor Statistics for the purpose of determining the feasibility of incorporating scanner data into the U.S. CPI; see Richardson (2000).

Artificial "populations" were drawn from these data, as we shall presently describe. Thus the study encompasses an apparently narrow world, that of cereal, within a fairly restricted geographic domain. Even this restricted world, however, allows for rather discrepant price trends over the six years. Thus, although we will not be able to generalize, in any simple fashion, to global price indexes encompassing a wide heterogeneity of products, we may be able to derive important clues on the behavior of different sampling methods and estimators.

The six years' worth of data available provided the opportunity for establishing fairly long price trends. In order to keep the data manageable and avoid the complications of seasonality, we limited ourselves to February data. For each February, for each item (i.e., each particular combination of brand, type, size) in a particular outlet, four weeks of price and quantity data were combined into a single month's price and quantity, by using the sum of quantities sold during the month as the quantity, and the *unit value* as the price. The unit value is an average price, defined as the sum of prices times quantities divided by the sum of the quantities. Unit values computed over short periods of time (e.g., a month) give perhaps the most meaningful sense of the "average" price for a particular item. The use of unit values smoothes the data and reduces it to more manageable proportions..

For the purposes of the study, the population of breakfast cereals was divided into four groups: (1) Hot Cereals (*H*), (2) "Sugary" cereals (*S*)– those in which some sweetening is a conspicuous component, (3) "Fruity" cereals (*F*) – those intended for the health conscious, and (4) "Ready-to-eat cereals" (*R*) –cold cereals not falling into categories (2) and (3). For each group, for each successive pair of years (using item-outlet combinations available in both years), superlative and non-superlative indexes were calculated. Long range indexes (95 to 00) were calculated both directly and by chaining. Additionally, indexes were calculated on the "core" items, those available in all six years.

Numerical results of the study will be made available elsewhere. Here we summarize our observations on the population indexes: (1) The superlative indexes differed relatively little from each other, a noteworthy result given the amount of variability in the item-outlet price relatives and quantities, due to "sales." (2) The non-superlative indexes differed wildly from each other and the superlatives. (3) Chained and direct superlative indexes did not differ much. (4) Indexes based on just the core items differed little from indexes using all relevant year-to-year data. (5) The price trends of the four major groups were quite different: *H* increased, *S* decreased sharply, *F* decreased modestly, *R* increased modestly.

Based on this preliminary investigation, we took the "populations of interest" to be 12 populations consisting of the four groups in each of the three regions, restricted to the core data. We took the target population parameters to be the year-to-year and long-range superlative indexes for each of these populations. For sampling purposes, the groups were divided into smaller groups. We call these "composite items" following the terminology of the International Labour Office –the ILO-- (see Turvey 1989), or "item strata" following the U.S. convention. This is the lowest level at which indexes are calculated. These groups were further divided into yet smaller groups of items (the "representative items" (ILO) or "entry level items" (U.S.). Grouping was based on rough anticipated similarity of items with respect to price trend. Data on cereal expenditures from the U.S. Consumer Expenditure Survey for the three regions in the relevant years were conjoined to the scanner data to construct artificial populations of households, whose expenditure patterns were realistic and added up to the total expenditures (by category) from the scanner data.

Then, through simulation, we compared the results of combining various modes of household and outlet sampling and using different index estimators. A consumer expenditure survey based on a simple random sample of households in each region may, for example, be combined with a price survey based on a purposive sample of items from a stratified sample of outlets, and a geomean index estimator may be applied at the item stratum level. In addition to classical methods in use in many countries, we investigated some new approaches, recently adopted or still under development, for example, the use of scanner data itself as an input to index estimation or the use of unit-values for item categories in which the individual items differ somewhat from period to period.

Index values and other characteristics of the scanner data, details of the household population construction, a description of the sampling methods and estimators investigated, and numerical results of the simulation study are presented in the Price/Production Indices Invited Paper Meeting of this 53rd Session of the ISI. They will be available afterwards in a longer version of this paper in the conference proceedings of the IASS, and on line at **http://stats.bls.gov/ore/pdf/st010020.pdf**.

REFERENCES

De Haan, J. , Opperdoes, E., and Schut, C. (1999), "Item Selection in the Consumer Price Index: Cut-off Versus Probability Sampling, *Survey Methodology*, 25, 1, 31-41

Richardson, D. H. (2000), "Scanner Indexes for the CPI," *Proceedings of the Conference on Scanner Data and Price Indexes*, NBER, Cambridge, http://www.nber.org/books/

Sarndal, C.-E., Swensson, B. and Wretman, J. (1992), *Model assisted Survey Sampling*, Springer, New York

Turvey, R. (1989), *Consumer Price Indices: An ILO Manual*, International Labour Office, Geneva

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000) *Finite Population Sampling and Inference*, Wiley, New York

RESUME

A partir de données fournies par scanner sur les prix et les quantités de céréales achetées, nous construisons plusieurs populations dont on connaî t les index du prix du consommateur. Ensuite, à travers la simulation, on compare la précision de l'estimation des populations cibles obtenue par une variété de méthodes d'échantillonnage courantes avec celle des estimateurs d'index différents. Ce qui relève d'un intérêt particulier, c'est la comparaison de la probabilité avec des approches intentionnelles à l'échantillonnage.