SAMPLE REDESIGN FOR THE INTRODUCTION OF THE TELELPHONE POINT OF PURCHASE SURVEY FRAMES IN THE COMMODITIES AND SERVICES COMPONENT OF THE U.S. CONSUMER PRICE INDEX

Sylvia G. Leaver, William H. Johnson, Owen J. Shoemaker, Thomas S. Benson
United States Bureau of Labor Statistics, 2 Massachusetts Avenue, N.E., Rm. 3655, Washington, D.C. 20212

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

KEY WORDS: Sample design; Optimization; Components of Variance

This paper describes the methods used to allocate data collection resources for the most recent redesign of the sample for the commodity and services (C&S) component of the U.S. Consumer Price Index. These methods rely on models relating price change sampling variance and data collection costs to design variables which are the number of items to price and outlets to visit per item group in each sample city. With these models, the optimal allocation of data collection resources to minimize sampling variance of price change, subject to budgetary and operational constraints, can be found using nonlinear programming techniques. This work represents an expansion of models developed for the 1987 and 1996 C&S sample redesigns. Models for sampling variance and costs are given, and solutions to the design problem posed under varying assumptions are discussed. A closing section characterizes the changes in sample allocation from previous designs.

Background

For a full discussion of the Consumer Price Index (CPI), we refer the reader to Chapter 17 of *The BLS Handbook of Methods* (1997). See also Leaver and Valliant (1995) for a more detailed description of the C&S sample design, selection, and estimation procedures

The current CPI is a Laspeyres aggregation of a combination of Laspeyres- and geometrically-averaged sub-indexes computed for the breadth of consumer-purchased commodities and services. The C&S component, which represents 72.5% of the expenditure weight of the CPI, is computed from measurements of price change on a sample of commodities and services, collected from selected outlets in sample cities across the United States. Consumer items are grouped into strata, the most finely defined item classes for which a price index is computed. Let IX(*I,t,0*) denote the month t index for a collection of strata, termed an item aggregate *I*, where month *0* represents the index base or reference period. Then

$$IX(I,t,0) = 100\left(\left(CW_t(I)\right)/\left(CW_0(I)\right)\right)$$

$$= \left(\sum_{i \in I} RI_t(i)\, IX_t(i)\right) \Big/ \left(\sum_{i \in I} RI_0(i)\, IX_t(i)\right)$$

where $CW_t(I)$, is an estimate of expenditures (called a cost weight) on a collection of items during period *t*, computed as a weighted average of individual indexes over all item strata in the item aggregate *I*, where $RI_t(i)$ is the expenditure-weighted relative importance of item *i* at time *t*.

In this application, we were concerned with the short term or *d*-month percentage price change:

$$PC\ (I, t, t-d) = 100 \cdot \left[\frac{IX\ (I, t, 0)}{IX\ (I, t-d)} - 1 \right]$$

An index area is the most basic geographic area for which a price index is computed on a monthly, bimonthly, or semiannual basis. There are two types of index areas: self-representing areas, such as New York, which were selected with certainty; and non-self-representing areas, whose sample comprises two or more primary sampling units (PSU's) selected according to a probability sample. The 1998 revised U.S. All Cities CPI is a weighted average of 38 index area CPI's; 31 from self-representing and 7 from non-self-representing areas. For purposes of variance estimation and operational manageability, the sample for each index area is segmented into two or more subsets, called replicate panels.

Each item stratum is composed of one or more narrowly defined classes called entry level items (ELI's). An ELI describes the level of specification for a class of goods with which a data collector enters an outlet for initial pricing.

In CPI sample selection, ELIs are selected from each stratum by a systematic probability proportional to size (pps) procedure, where, with the 1998 revision, the ELI weights were derived from expenditures reported in the 1993-1995 Consumer Expenditure Surveys. ELI selections are independently drawn for each replicate panel within each PSU.

Sample frames and weights used in outlet selection are derived from the Telephone Point of Purchase Survey (TPOPS), a random digit telephone survey conducted by the U.S. Bureau of the Census for the

BLS. Beginning in 1997 the TPOPS replaced the Continuing Point of Purchase Survey, a household survey. The TPOPS survey provides the names and addresses of outlets and dollar amounts of purchases, for item classes known as POPS categories. A POPS category is a class of items which are normally sold in the same kind of outlet. Each ELI belongs to only one POPS category. Outlet frames and selection weights are derived from POPS survey data for each PSU-POPS category-replicate panel.

In outlet selection, outlets are selected by systematic pps from frames for each PSU-replicate panel for POPS categories corresponding to ELIs selected in item sampling. Selected items are then priced in sample outlets on a monthly, bimonthly, or seasonal basis.

History

Hansen, Hurwitz,and Madow (1953), Kish (1965), and Cochran (1977) present several examples of sample design optimization via cost and error modeling. Groves (1990) discusses sample design for social surveys.

Cost and sampling error models were first formulated for the C&S sample design for the 1978 CPI Revision (Westat, 1974). Item classes comprised two categories - food, and other goods and services, and sample size allocation were made for six PSU classes. Selection of the sample design implemented in that revision was based on evaluation of a number of alternative designs. The 1987 CPI Revision (CPIR) redesign (Leaver, et al., 1987) expanded on this approach, refining models for eight item groups and ten PSU classes. This implementation relied on detailed use of administrative records and modeled estimates for cost and variance function estimates. Solution methods used nonlinear programming techniques to identify local minimizers of a modeled relative variance function, under varying assumptions of annual inflation and price change interval. For another BLS survey, Valliant and Gentle (1994) developed a generalized system for constrained optimization of a two-stage stratified sample design implemented on a UNIX platform, with a weighted summed relative variance objective function.

The approach taken in this application generally follows that taken for the 1987 and 1998 CPIRs (Leaver, et al., 1996). Sampling variance was minimized in this application. Data collection cost models were revised; costs were derived from administrative records and a time and travel study of CPI data collection. The size of the nonlinear programming problem solved was expanded, and detailed distribution of item-outlet sampling resources used stratum-level variance estimates not previously available for sample design allocation.

The Design Problem

The primary objective of the C&S sample redesign was to determine values for all sample design variables which would minimize the sampling variance of price change for the C&S portion of the CPI. Sample design variables for the C&S component were the number of ELI's to select in each item stratum and the number of outlets to select per CPOPS category-replicate panel in each sample PSU. The number of PSU's, the number of replicate panels per PSU, and the item stratification were previously determined (Williams et al., 1993; Lane, 1996 and Williams, 1996.)

Certain simplifying assumptions were made to render the problem tractable. Newly revised item strata were divided into thirteen item groups: four subgroupings of food at home, food away from home and alcoholic beverages, household furnishings and operations, fuels and utilities, apparel, transportation less motor fuel, motor fuel, medical care, education and communications, and the combined group of recreation and other commodities and services. The 87 PSU's were divided into 15 groups according to size and number of replicate panels. It was assumed that the same outlet sample sizes would apply to all PSU's within the same PSU group. It was also assumed that the same item sample selection sizes would apply across all PSU's . This reduced the allocation problem to one of determining the values of the design variables $\{ K_i , i=1,...,13\}$, the number of ELI selections per item group-replicate panel within each PSU and $\{ M_{ij} , i=1,...,13, j =1,...,15 \}$, the designated number of outlet selections per item group-POPS category-replicate within each PSU, which would minimize a modeled price change sampling variance, subject to additional allocation and cost constraints.

The variance of price change for all C&S items was modeled as a function of the design variables, as were total annual data collection costs. Nonlinear programming methods were then used to determine optimal values for the design values under various cost, variance, and sample share constraints. Detailed descriptions of these activities follow.

The Sampling Variance Function

For the purposes of the allocation problem, we write the All U.S. City Average C&S price change estimator as $PC(\cdot,\cdot,t,t-d) = \sum_i \sum_k RI_{i,k} PC(i,k,t,t-d)$, where $PC(i,k,t,t-d)$ is the estimated price change from time t-d to

t for item group *i* and index area *k*, and $RI_{i,k}$ is the population-expenditure-weighted relative importance of item group *i* in index area *k*. Deriving a component form of the variance of this price change estimator, accounting for the stages of sampling described above, would be extremely difficult. Rather than this direct route, we have taken a more indirect, modeling approach described below. Four sources of variation were modeled: PSU selection, item selection, outlet selection, and other sources, such as sampling within the outlet.

The variance function for the CPI revision was modeled for index areas. Each self-representing PSU is a single index area. Non-self-representing PSU's were selected to represent 7 index areas, whose sample consisted of 2 to 22 PSU's. The variance model assumes that the total variance of price change for item group *i* within index area *k* can be expressed as a sum of four components:

$$s^2_{i,k} = s^2_{psu,i,k} + s^2_{eli,i,k} + s^2_{outlet,i,k} + s^2_{error,i,k}$$

where

$s^2_{i,k}$	is the total variance of price change for item group *i* in index area *k*,
$s^2_{psu,i,k}$	is the component of variance due to sampling PSU's in non-self-representing areas, 0 for self-representing areas,
$s^2_{eli,i,k}$	is the component of variance due to sampling of ELI's within item strata,
$s^2_{outlet,i,k}$	is the component of variance due to sampling of outlets, and
$s^2_{error,i,k}$	is a residual component of variance attributable to other aspects of the sampling process, including the final stage of within-outlet item selection, called disaggregation .

We assume that the variance of price change of an individual sampled unit or quote has the same structure:

$$s^2_{unit,i,k} = s^2_{unit,psu,i,k} + s^2_{unit,eli,i,k} + s^2_{unit,outlet,i,k} + s^2_{unit,error,i,k}$$, where

$s^2_{unit,i,k}$	is the total variance of price change of an individual sampled unit or quote for item *i* in area *k*,
$s^2_{unit,psu,i,k}$	is the component of unit variance due to sampling PSU's in non-self-representing areas,
$s^2_{unit,eli,i,k}$	is the component of unit variance due to sampling of ELI's within item strata,
$s^2_{unit,outlet,i,k}$	is the component of unit variance due to sampling of outlets, and
$s^2_{unit,,i,k}$	is the corresponding residual component of unit variance.

It follows that each component of $s^2_{i,k}$ can be written in terms of its corresponding unit variance components:

$$s^2_{i,k} = s^2_{unit,psu,i,k}/N'_k + (s^2_{unit,item,i,k}/(N_k H_k K_i))NC_i$$

$$+ s^2_{unit,outlet,i,k}/(N_k H_k M'_{i,k}) + s^2_{unit,error,i,k}/(N_k H_k K_i M_{i,k})$$

where

N_k	is the number of PSU's in index area *k* ,
N'_k	is the number of non-self-representing PSU's in the index area,
H_k	is the number of replicate panels per PSU in the index area,
$M'_{i,k}$	is the number of unique in-scope outlets selected per PSU-replicate
NC_i	is the percent of strata in item group *i* which are non-certainty strata.

Note that the expected number of quotes per PSU-replicate panel- item group is estimated by the product of the designated outlet sample size and the number of item stratum selections, $M_{ij} \cdot K_i$.

Thus the sampling variance of price change for the All U.S. City Average C&S index is

$$s^2_{TOTAL} = \sum_k \sum_i RI^2_{i,k} s^2_{i,k}$$

The Cost Function

The total annual cost of the C&S portion of the CPI includes costs of initiation data collection and processing, personal visit and telephone pricing, and pricing data processing, each of which were developed in terms of outlet and quote related costs. For PSU group *j* and item group *i*, outlet related costs for initiation are:

$$CI_O(M_{ij},K_i) = 0.25N_j \cdot H_j \cdot C_{O,i} \cdot (a_{ij}M^2_{ij} + b_{ij}M_{ij})$$, where

$CI_O(M_{ij},K_i)$	is the outlet-related initiation cost for item group *i* in PSU group *j*
N_j	is the number of PSU's in group *j*,
H_j	is the number of replicates per PSU in PSU group *j*,
$C_{O,i}$	is the initiation cost per outlet for item group *i*,

and $(a_{ij}M_{ij}^2 + b_{ij}M_{ij})$ is an overlap function used to predict the number of unique sample outlets, accounting for the overlap of elements in the outlet sample within and between item groups for a replicate panel. The number 0.25 accounts for the rotation or reselection of one-fourth of the sample each year.

Quote related initiation costs are:

$$CI_Q(M_{ij}, K_i) = 0.25 N_j H_j \cdot SeasI_i \cdot C_{Q,i} \cdot M_{ij} \cdot K_i \cdot NR_i$$

where

$CI_Q(M_{ij}, K_i)$	is the quote-related cost of initiation for item group *i* in PSU group *j*,
$SeasI_i$	is a seasonal items initiation factor for item group *I*,
$C_{Q,i}$	is the initiation cost per quote for item group *i*, and
NR_i	is the outlet initiation response rate for item group *i*.

The costs of ongoing price data collection and processing were also developed as both outlet and quote related costs. For PSU group *j* and item group *i*, outlet related costs for ongoing pricing are:

$$CP_O(M_{ij}, K_i) = MB_{ij} \cdot N_j \cdot H_j \cdot NR_i \cdot (a_{ij}M_{ij}^2 + b_{ij}M_{ij}) \cdot [(C_{PV,O,i} + C_{PV,T,i}) \cdot (1 - R_{T,O,i}) + C_{T,O,i} \cdot R_{T,O,i}]$$

where

$CP_O(M_{ij}, K_i)$	is the total outlet-related cost for ongoing pricing,
$C_{PV,O,i}$	is the cost for a personal visit for pricing per outlet for item group *i*,
$C_{PV,T,i}$	is the travel cost for a personal visit for pricing per outlet for item group *i*,
$R_{T,O,i}$	is the proportion of outlets priced by telephone for item group *i*,
$C_{T,O,i}$	is the per outlet cost for telephone collection,
MB_{ij}	is a factor to adjust for the monthly/ bimonthly mix of outlets and quotes by PSU and major product group.

Quote related costs for ongoing pricing are:

$$CP_Q(M_{ij}, K_i) = MB_{ij} \cdot N_j \cdot H_j \cdot M_{ij} \cdot K_i \cdot NRQ_i \; SeasR_i \cdot [C_{PV,Q,i} \cdot (1 - R_{T,Q,i}) + C_{T,Q,i} \cdot R_{T,Q,i}]$$

, where

$CP_Q(M_{ij}, K_i)$	is the total quote-related cost for ongoing pricing,
$C_{PV,Q,i}$	is the per quote cost for a personal visit for pricing,
$R_{T,Q,i}$	is the proportion of telephone collected quotes for item group *i*,
$C_{T,Q,i}$	is the per quote cost for telephone collection for item group *i*, and
NRQ_i	is the quote level pricing response rate for item group *i*.
$SeasR_i$	is a seasonal items ongoing pricing factor for item group *i*.

The total cost function associated with data collection and processing for C&S, summed over all item groups and PSU groups, is then given by:

$$C_{Total} = \sum_{i,j} [CI_O(M_{ij}, K_i) + CI_Q(M_{ij}, K_i) + CP_O(M_{ij}, K_i) + CP_Q(M_{ij}, K_i)]$$

Thus, the sample design problem can be expressed as the nonlinear programming problem:

Minimize $s^2_{Total}(\{K_i\}, \{M_{ij}\})$ subject to:

$C_{Total} \le \$5{,}300{,}000$

$K_i \ge$ *Number of item strata in item group i,*

$K_i \le$ *Maximum number of item hits for item group i,*

$M_{ij} \ge 2$, *i=1,..., 13, j=1,...,15*

Average item hits per stratum ≥ 9

Model coefficients

Estimates of components of the cost function were developed using agency administrative records. Fiscal year 1996 data were used to obtain a total cost per outlet to initiate, and then data provided by the field office produced a per hour cost of initiation. Outlet unit costs and quote unit costs of initiation, by item group, were derived by taking these per outlet and per hour costs and combining them with data obtained from a data collection time and travel study conducted in 1987. Travel costs per quote, by item group, were estimated by using an overall travel cost per outlet and again comparing it to data from the 1987 time and travel study.

Pricing costs were figured in a similar manner. Distinctions between personal visit and telephone collection of data were made based upon cost accounting information from the field office and from an analysis conducted within the Prices Statistical Methods Division. Outlet initiation survival rates and quote and outlet retention rates for each item group were developed from field initiation records and ongoing pricing records for mid 1993-mid 1997.

"Overlap" functions were modeled to project the number of unique outlets realized in sample selection as a function of designated sample size. These were obtained by modeling the number of unique outlets

obtained in simulations of sampling procedures for each PSU and item group, using CPOPS and TPOPS sampling frames for the most recent rotations for each PSU-item stratum (Johnson et al., 1999).

Components of price change variance were computed using weighted restricted maximum likelihood components of variance estimation methods and C&S price micro-data collected from 1993-97 (Shoemaker and Johnson, 1999). Component estimates were developed for 6-, and 12-month price change for the 13 item groups for each index area.

Problem Solution

SAS NLP was used to find a local minimum to the design problem. Solutions were found using components of variance estimates for 6-, and 12-month price change components of variance estimates. For each item group, the number of item selections was bounded below by the number of strata in the item group and above by a ceiling of 133% of the item group's pre-1998 revision item stratum hits allocation.

Only minor differences were observed between the problem solutions found for differing pricing intervals. The solution found for 6-month price change was selected because variance component estimates were considered most stable for this interval.

Item hits were then distributed among item strata within each item group, with consideration given to differences in relative importance, stratum level price change variance estimates, and response rates among the item strata within each item group, as well as special problems identified by commodity analysts and field staff. Similarly, designated outlet sample sizes were adjusted among the various POPS categories in item groups to manage variation in expected response rates and respondent burden.

Although major revisions in the CPI occur every 10 years, incremental revisions are planned with each year with the staggered rotation of TPOPS categories in sample PSUs. The table below characterizes the latest TPOPS rotation sample design, contrasting it with the design implemented in sample rotations for the four years prior to the last CPOPS rotation revision, which occurred in 1996. In general, the sample design represents a considerable reduction in modeled sampling variance from that projected under the same model for these prior allocations. This reduction is primarily attributable to a 33% increase in the data collection resource budget; the gains in projected precision match almost one to one with gains in allocable budget. In addition, the allocation also shifted resources in many item groups from sampling many outlets to fewer outlets, with more item selections per outlet. This is due primarily to the larger residual component of price change sampling variance estimated for most item groups. This component was regarded as negligible in earlier estimation (Leaver, et al., 1987).

Acknowledgments

The authors would like to thank Janice Lent, David Chapman, and Alan Dorfman for their careful reading of earlier drafts of this paper and their helpful comments, Glenn Pontanilla and Jane Martinez of the Office of Field Operations, and Valerie Harris of the Office of Prices and Living Conditions for their assistance in survey cost accounting, and Kirk Hagemeier and Sean McIllece of the Prices Statistical Methods Division, Robert Cage of the Division of Consumer Prices and Price Indexes, and Mark Crosby of the Division of Consumer Prices and Consumption Studies for their assistance in obtaining production sampling frames. The authors also would like to thank Janet Williams and Brian Hedges for their support on this project.

References

Bureau of Labor Statistics (1997) *BLS Handbook of Methods*, Washington. DC: U.S. Government Printing Office, pp. 167-230.

Cochran, W. G. (1977). *Sampling Techniques*, Wiley, New York.

Groves, Robert (1990) *Cost and Error Modeling in Social Science Surveys*, Wiley, New York.

Hansen, Morris G., Hurwitz, William N., and Madow, William G. (1953) *Sample Survey Methods and Theory*, Wiley, New York.

Johnson, William H., Leaver, S.G., and Benson, T.S. (1999) " Modeling the Realized Outlet Sample for the Commodities and Services Component of the U.S. Consumer Price Index," *Proceedings of the Government Statistics Section, American Statistical Association*, to appear.

Kish, Leslie (1965) *Survey Sampling*, Wiley, New York.

Lane, Walter (1996) "Changing the CPI Item Structure," U.S. Bureau of Labor Statistics http://www.stats.bls.gov/mlr/cpiwl001.htm

Leaver, S., Weber, W., Cohen, M., and Archer, K. (1987) "Item-Outlet Sample Redesign for the 1987 U.S.

Consumer Price Index Revision," *Proceedings of the*

Comparison of Modeled Sampling Error and the Distribution of Sample Resources between Pre-revision and Latest Allocation of C&S Design

Item Group	Root Mean 6-Month PC Variance 9807-9901	% Change in Modeled PC SE from 1994 Rotation to Latest Revision	% Total Costs, 1990-94 Rotation	% Total Costs, 1999 TPOPS Design	% Total Quotes, 1999 TPOPS Design	92-94 CE Relative Importance (relative to total C&S)
Total, All Items less Rent and Owners' Equivalent Rent	.0977	-15.9	100.0	100.0	100.0	100.0
Food at home (4 groups)	.2312 .3430 .6099 .1956	-2.6 +1.5 -2.7 +0.6	5.3 4.8 4.6 13.2	3.3 2.9 2.3 7.1 Total:15.6	6.8 8.5 6.5 8.8	3.38 3.41 1.78 5.18 Total: 13.75
Food away + Alcoholic Beverages	.1201	-2.0	5.3	4.8	5.4	8.14
Household Furnishings & Operations	.4992	-26.4	8.1	12.3	12.1	10.41
Fuels and Utilities	.3524	-2.0	4.4	4.2	5.3	6.91
Apparel & Upkeep	.8844	-11.1	13.5	11.7	8.2	7.34
Transportation less Motor Fuels	.1435	-19.1	14.4	20.33	13.4	20.01
Motor Fuels	.2864	-26.7	3.1	2.72	2.9	4.23
Medical Care	.1985	-21.3	6.0	6.23	5.0	7.35
Education & Communication	.2364	-17.9	5.1	7.15	5.7	7.58
Recreation + Other C&S	.1899	-19.7	12.3	15.22	11.4	14.26

46th Session, International Statistical Institute. Tokyo, Vol. 3, pp. 173-185.

Leaver, S., Johnson, W. H., Baskin, R., Scarlett, S., and Morse, R. (1996) "Commodities and Services Sample Redesign for the 1998 Consumer Price Index Revision," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Vol. I, pp. 239-244.

Leaver, S. and Valliant, R. (1995) "Chapter 28: Statistical Problems in Estimating the U.S. Consumer Price Index," *Business Survey Methods*, Brenda G. Cox, et al., editors, Wiley, New York.

Shoemaker, O., and Johnson, W. H. (1999) "Estimation of Variance Components for the U.S. Consumer Price Index", *Proceedings of the Survey Research Methods Section, American Statistical Association,* to appear.

Valliant, R., and Gentle, J. (1994), "An Application Of Mathematical Programming to Sample Allocation," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 683-688.

Westat, Inc. (1974). "Proposals for and Evaluation of Alternative Designs for Allocation of CPI Pricing Efforts to Items, Outlets, and within Outlets," CPIR-WS-4, Rockville, Maryland.

Williams, J.L., Brown, E.F., Zion, G.R. (1993), "The Challenge of Redesigning the Consumer Price Index Area Sample," *Proceedings of the Survey Research Methods Section, American Statistical Association* (Vol. 1), pp. 200-205.

Williams, Janet L. (1996), "The Redesign of the CPI Geographic Sample," U.S. Bureau of Labor Statistics http://www.stats.bls.gov/mlr/cpijw001.htm