A SYSTEM OF VARIANCE ESTIMATION FOR THE U. S. CONSUMER PRICE INDEX

William L. Weber, Bureau of Labor Statistics

Introduction

The U.S. Consumer Price Index compiled and published by the Bureau of Labor Statistics (BLS) is one of the nations primary measures of inflation. It provides critical information for policy decisions by business, labor, and government leaders. The CPI is also widely used as an escalator of wages, pensions, alimony and other contractual payments. The current CPI was extensively revised during the 1970's. This revision was the most complex and extensive revision of the CPI since its origination during World War I. Virtually every aspect of the CPI methodology was reviewed, analyzed and improved as part of the revision effort. With the publication of the revised CPI in February, 1978, the question of the variance of the new index was raised.

The BLS is now implementing a system of variance estimation for the CPI. The subject of this paper is that variance estimation system. Estimates of variance will give the CPI data users additional decision making information. These estimates will also provide the BLS with a valuable tool for use in improving and enhancing the CPI. Before discussing the variance estimation methodology, we will describe the CPI sample design. Two CPI indexes are actually produced each month: One for the urban wage and clerical population, corresponding to the unrevised CPI, and one for the all urban population. Ideally, each CPI would have as inputs the total set of prices, placed in one-to-one correspondence to the total set of purchases of all members of the index family population. However, the magnitude of these variables and the constraints of time and cost will not permit the achievement of this ideal. Therefore a complex, multi-stage sample design is employed to produce the monthly CPI.

Definition of the U.S. Consumer Price Index[1/]

The U.S. CPI is a Laspeyres index of price change from time 0 to time t. The Laspeyres index reflects only changes in price for a constant quality set of items. It can be expressed as:

$$I_{zt,o} = \frac{P_{zt}Q_{zo}}{P_{zo}Q_{zo}} \times 100.0$$

where Q_{zo} is the quantity of item z purchased in the base period (time zero), P_{zt} is the average price for item z at time t, and P_{zo} is the average price for item z in the base period. The product, $P_{zo}Q_{zo}$, is the amount of total expenditures for item z in the base period. That is, $P_{zo}Q_{zo}$ is the base period cost weight for item z. Thus, the CPI for item z is the ratio of the current period cost weight to the base period cost weight. The index can be rewritten as:

$$I_{zt,o} = \frac{P_{zt-1}Q_{zo}\,\frac{P_{zt}Q_{zo}}{P_{zt-1}Q_{zo}}}{P_{zo}Q_{zo}} \times 100, \text{ where}$$

$\frac{P_{zt}Q_{zo}}{P_{zt-1}Q_{zo}}$ is a measure of a one period price change for item z. This equation illustrates that the estimation of the U.S. CPI is a two-fold problem. First, the base period cost weights $(P_{zo}Q_{zo})$must be estimated. And second, the one period price change, $\frac{P_{zt}Q_{zo}}{P_{zt-1}Q_{zo}}$ must be estimated for each pricing period.

The base period cost weights are estimated from data collected in the Consumer Expenditure Survey by the U.S. Bureau of the Census for BLS. The one period price change or price relative, denoted by $R_{zt,t-1}$, for all CPI commodities, services, rent and property tax is estimated from data collected each pricing period by BLS.

The price relatives for the house purchase and mortgage interest components are estimated from data provided by the Federal Housing Administration (FHA). All of these activities are described in succeeding sections of this paper.

The Consumer Expenditure Survey

The source of the base period cost weights currently employed in the CPI is the 1972-73 Consumer Expenditure Survey (CES) conducted by the Census Bureau for BLS. The Census Bureau began data collection in September, 1979 for a Continuing Consumer Expenditure Survey (CCES). Data from the (CCES) will be used to update the cost weights. Since this data will not be introduced into the CPI for some time, only the CES will be described here. The CCES parallels the CES in its basic design.

The CES consisted of two separate components each with its own questionnaire and sample. The first component was a diary survey completed by respondents for two consecutive one week periods. The objective of the diary survey was to obtain reliable expenditure data on small frequently purcashed items which are normally difficult to recall. These items include expenditures for food and beverages, natural gas and electricity, gasoline, housekeeping supplies, non-prescription drugs and medical supplies, and personal care products and services. Consumer units were asked to list all expenses for these items during the survey period. Data on income and family characteristics was also collected.

The second component of the CES, called the quarterly survey, was an interview panel survey in which each consumer unit in the sample was visited by an interviewer every three months over a fifteen month period. This survey was designed to collect information on major items of expense as well as on income and family characteristics. Items reported on the quarterly included expenditures for the following: housing, household equipment, house furnishings, vehicles, subscriptions, insurance, educational expenses, clothing, repair and maintenance of property, utilities, fuels, vehicle operating expenses and expenses for out of town trips. The final interview in the fifth quarter provided the regularly recorded expenses plus information on homeownership costs, work experience, changes in assets and liabilities, estimates of consumer unit income and other selected financial information.

The samples of consumer units for the CES were selected as follows. For both the diary and quarterly survey the nation was stratified into 216 geographic strata using stratification variables defined for the Current Population Survey of the Census Bureau. Thirty of these areas were designated as self-representing. Half of the housing units in each self-representing area was covered in the first survey year and half in the second survey year. The 186 NSR areas were divided into two 93-area groups. Each of these groups was covered in one of the two survey years. A primary sampling unit was randomly selected from each of the 186 areas using a controlled selection procedure to insure the desired geographic distribution.

Housing units within the 216 geographic strata were stratified by income level, housing tenure, and size of primary family. The actual sample of unclustered housing units for the diary survey was selected systematically by computer from the 1970 Census 20 percent sample data file. About 30,000 housing units were selected for diary interviews for each year of the survey. The sample of housing units for the quarterly survey was also selected by computer from the 1970 Census 20 percent sample data file. Approximately 20,000 housing units were designated for interview over the two years of the quarterly. For both surveys, sample housing units within a PSU were distributed by month to permit data collection throughout the survey year.

The base period cost weights ($P_{zo}Q_{zo}$) were estimated as follows. Weights were assigned to each consumer unit in the survey. Each weight was the product of the following factors: the inverse of the probability of selection of each housing unit, non-interview adjustment factor, in non-self-representing PSU's a ratio estimation factor for color and residence, a ratio estimation factor for age, sex, and color as derived from known civilian population controls. An adjustment for family composition was also made. The weights were applied to the expenditures reported by each consumer unit. The weighted expenditures were then applied to the CPI item strata level.

CPI Item Sample

The CES also provides the sampling frame used to select the general categories of items priced for the CPI. The basic structure for aggregating the diary and quarterly expenditure data serves as the item sampling frame. This structure consists of 68 expenditure classes (EC's) which correspond to the primary publication levels of the indexes. Within each EC the expenditures are grouped into one or more item strata. There are a total of 265 item strata. Within each item strata, one or more substrata, called Entry Level Items (ELI's) are defined. There are a total of 382 ELI's. ELI's are the ultimate sampling unit selected in Washington. They are used by the data collectors as their initial level of item definition within an outlet.

Four regional market basket universes were tabulated from the diary and quarterly surveys to reflect any regional differences. The diary provided the expenditures for frequently purchased items such as food and personal care items. The quarterly provided expenditure data for all other items. Within each of the four regions eight independent samples of ELI's were selected for each item stratum. Thus, eight sample replicates of ELI's were formed for each region -- thirty-two sample replicates nationally. Each CPI PSU replicate was assigned one of the eight item sample replicates from the corresponding region for pricing.

Because of the requirement of publishing indexes for two family definitions, the following technique was used to select the item samples. The ELI sample was selected first for the urban wage and clerical population (W). This sample was selected systematically with each ELI assigned a probability proportional to the expenditures reported on the CES by the W population. The ELI sample for the all urban population (U) was then selected using a Keyfitzing technique to maximize the overlap of sample ELI's between populations and maintain the correct probabilities of selection.

CPI Pricing Area Design

Price data for the CPI is collected in 85 geographic areas or primary sampling units (PSU's). These 85 areas were selected for pricing as follows. The entire country was divided into PSU's. Each Standard Metropolitan Statistical Area (SMSA) defined by the Census Bureau was defined to be a PSU. The counties not included in an SMSA were grouped into contiguous areas to form PSU's. The following variables were used to form 85 strata from the PSU's: the percent of population increase from 1960 to 1970; the degree of major industrialization; the percent of the population which was non-white; and the percent of the population which was urban. This design yielded 27 self-representing strata and 58 non-self-representing strata. One PSU was selected for pricing for each non-self-representing strata. A controlled selection procedure was employed for sampling the PSU's to insure that the sample areas were distributed geographically across the country.

CPI Outlet Sampling, Within Outlet Item Sampling

The sampling frame of outlets for all food items and most commodities and services in the CPI was constructed from data collected by the Census Bureau in the Point of Purchase Survey (POPS). The POPS survey was a household survey commissioned by BLS and conducted by the Census Bureau in the 85 PSU's. Expenditure data plus the name and address of the place of purchase were collected for about 100 relatively broad categories of items from each respondent. The original POPS survey was conducted during 1973-74 and included about 23,000 housing units. A Continuing Point of Purchase Survey (CPOPS) has been conducted in one-fifth of the PSU's each year.

The POPS and CPOPS do not provide a sampling frame of outlets for approximately 40% of the CPI items by expenditure weight. The items not covered by the POPS are grouped together under the heading Non-POPS. They include rent, property tax, mortgage interest, house prices, utilities, transportation, insurance and several miscellaneous categories. Some of these items were excluded from the POPS because existing sampling frames were adequate. For others, it was felt that the POPS would not yield an adequate sampling frame. Thus, for each Non-POPS component the sampling frame was either purchased or acquired from another agency or constructed by BLS staff. Sample designs for the rent, property tax, and mortgage interest components are described in later sections of this paper.

The following procedures were employed for selecting the outlet samples from the POPS sampling frames. The total expenditures reported for each outlet were edited. Outlets with zero expenditures were given a minimum chance of selection. Outlets with very large expenditures were edited down to 20% of the total expenditure for the U population for the PSU/POPS category. A systematic PPS sample of outlets was selected for each POPS category for the W population. The measure of size associated with each outlet was proportional to the average daily expenditure reported for the outlet by all households of the W population. A Keyfitzing technique was then used to recompute the measures of size for each outlet in the universe for the U population. The outlet samples for each POPS category for the U population were then selected systematically with each outlet assigned a probability of selection proportional to the Keyfitzed measure of size.

Outlet samples for each of the Non-POPS commodities and services items were also selected systematically. Each outlet on a given frame was assigned a probability of selection proportional to its measure of size. Since a single frame covered both populations for each Non-POPS item, only one sample was selected for each item. The single sample was priced for both population indexes.

For each food, commodity or services ELI selected for pricing, a specific store item was selected in each corresponding sample outlet using a multi-stage probability sampling technique. Under this procedure each item in the ELI sold by the sample outlet is given a probability of selection proportional to the dollar sales of that item in the given outlet. Use of this procedure permits a probability sampling of outlets and quotes throughout the CPI.

Price Relatives for Food, Commodities and Services

At the end of each pricing period, the estimate of the one period (t-1 to t) price change (price relative) is computed for each item stratum and population. Only price quotes obtained in both the current and previous price period for the same or comparable item are used in the estimate. The same quote weights are used for both the current and previous period price quotes. The estimate of the one period price change for the zth item stratum for a given market basket is computed as:

$$R_{z,t,t-1}=\frac{R_{z,t,0}}{R_{z,t-1,0}}=\frac{\sum\limits_{h}\sum\limits_{i=1}^{n} W_{hijt}(P_{zijt}/P_{zij0})\Big/\sum\limits_{h}\sum\limits_{i=1}^{n} W_{hijt}}{\sum\limits_{h}\sum\limits_{i=1}^{n} W_{hijt}(P_{zijt-1}/P_{zij0})\Big/\sum\limits_{h}\sum\limits_{i=1}^{n} W_{hijt}}$$

where:

The subscript h denotes the possibility of summing across different PSU's in the B, C, D market baskets;

P_{zijt} is the price of the ith quote in the jth outlet in the current pricing period, t, for item z;

P_{zijt-1} is the price of the ith quote in the jth outlet in the previous pricing period, t-1, for item z;

P_{zij0} is the base period price for the ith quote in the jth outlet for item z;

W_{hijt} is the quote weight for the ith quote in the jth outlet in the current pricing period for item z.

The quote weight, W_{hijt}, consists of the product of the following values: an estimate of the total daily expenditure for the POPS category for the PSU and population (U or W); a duplication factor to reflect any special subsampling of outlets or quotes; the percent of sales of the ELI to the total sales of the POPS category in the jth outlet; the inverse of the proportion that the ELI is of the item stratum; the inverse of the number of useable quotes for the item stratum and PSU.

Rent and Property Tax Sample Design

A probability sample of housing units was selected in each of the 85 PSU's and screened for tenure. A subsample of renter occupied housing units is used to estimate the rent price relative and another subsample of owner occupied housing units is used to estimate the property tax price relative.

The CPI rent sample is divided into six panels. One panel is interviewed each month and a given panel is interviewed every six months. The current and previous months rent is collected for each on cycle sample unit. Thus, both one month and six month price changes are available. Each rent sample unit is assigned an internal weight which is the product of the following factors: the inverse of the unit's probability of selection; a duplication factor to reflect any special subsampling of units; and a non-interview adjustment factor.

The following process is used to estimate the price relative for rent for each CPI market basket. The weighted sums of all in-scope adjusted rents is computed for the current (t) and previous (t-1) months. The weights are the current months internal weights for each renter unit. Thus, we have

$a_{ht}=\sum\limits_{i=1}^{n_h} w_{hit}r_{hit}$ as the aggregate rent for market basket h in time t; and

$a_{ht-1}=\sum\limits_{i=1}^{n_h} w_{hit}r_{hit-1}$ as the aggregate rent for market basket h for time t-1. Similarly, the sums of the weighted rents is computed using those units with reported or imputed rents in t and t-6. We have

$$a^*_{ht}=\sum_{i=1}^{n_h} w_{hit}r_{hit} \quad\text{and}\quad a^*_{ht-6}=\sum_{i=1}^{n_h^*} w_{hit}r_{hit-6}.$$

One month and six month relatives are then computed using these aggregates:

$$r_{ht,t-1}=\frac{a_{ht}}{a_{ht-1}} \quad\text{and}\quad r_{ht,t-6}=\frac{a^*_{ht}}{a^*_{ht-1}}$$

These two relatives are combined using an artificial cost weight: $I_{ht} = WB_{ht-1}r_{ht,t-1} + (1-W)B_{ht-6}r_{ht,t-6}$

The final one period estimate of rent price change is then:

$$R_{ht,t-1} = \frac{I_{ht}}{I_{ht-1}}$$

Housing units included in the property tax sample may be taxed by more than one jurisdiction. Thus, the property tax data base contains a record for each housing unit/jurisdiction. The value of the property tax is computed separately for the current pricing period (T) and for the previous pricing period (T-1) for each housing unit/jurisdiction. The current survey period (T) is defined to include all property tax values obtained in months t through t-11. The previous survey period (T-1) is defined to include the property tax values obtained in months t-1 through t-12. The property tax value (GT) for the current period for a given housing unit/jurisdiction is computed as:

$G_T = \left((A_T - (\pm B_t) - F_T) \ X \ E_T\right) - L_T$ where

A = assessment
B = assessed dollar value of capital change
F = assessed dollar value of exemption
L = tax dollar value of exemption
E = tax rate

The property tax value (G_{T-1}) for the previous period is computed as:

$G_{T-1} = \left((A_{T-1} - F_{T-1}) \ X \ E_{T-1}\right) - L_{T-1}$

where the variables are as defined above.

The one month price relative for the hth market basket is defined as:

$$R_{ht} = \sum_{i=1}^{n} W_{hit} G_{Ti}(T,t-11) \qquad \sum_{i=1}^{n-6} W_{hit} G_{Ti}(t-1,t-11)$$

$$+ \sum_{i=1}^{n} W_{hit} G_{T-1i}(t-12)$$

where

W_{hit} = the weight for the ith housing unit in the hth PSU

n = the number of matched in scope housing unit/jurisdiction records in the hth PSU

n-b = the number of matched in-scope housing unit/jurisdiction records priced in periods t-1 through t-11

b = the number of matched in-scope housing unit/jurisdiction records priced in t and t-12

Mortgage Interest and House Prices:

The BLS does not directly collect data for either of these components. The data is collected by other federal agencies. For mortgage interest, the Federal Home Loan Bank Board provides a sample of mortgages each month from a sample of banks, savings and loans and mortgage companies. For the FHA/VA component of mortgage interest BLS uses the FHA/VA mortgage interest rate ceiling. An average mortgage interest rate is computed each month for each PSU and lender strata. A weighted average of the average interest rates is then computed for months t and t-1. The weights are estimates of the relative mortgage interest costs for 1972 - 1973. The final price relative is the product of the one-month price change for mortgage interest and the one month change in house prices.

Each month the Federal Housing Authority provides BLS with the universe of sales of FHA insured loans from the previous month. All non-FHA house sales are excluded from the measurement of price movement by the CPI. This limitation makes the house price component the weakest element of the CPI. The FHA housing units are stratified by age and living area. The average price per square foot is computed for each stratum for each market basket as:

$$\overline{PSF}_{jt} = \frac{1}{njt} \sum_{i=t}^{njt} PSF_{jit}$$

where n_{jt} is the number of houses in the jth stratum at time t; PSF_{jit} is the price per square foot for the ith house in the jth stratum at time t.

The price relative for each stratum for each market basket is then computed as:$R_{jt,t-1}=\overline{PSF}_{jt}/\overline{PSF}_{jt-1}$. The price relative for a given market basket is computed as: $R_{t,t-1} = (\Sigma V_{jt-1} \text{ X } R_{j,t,t-1})/\Sigma V_{j,t-1}$ where V_{jt-1} is a weight reflecting the total value of homes in stratum j at time t-1.

In summary, we can estimate the national CPI as:

$$I_t = \frac{\sum_h \sum_i BW_{hit-1} R_{hit,t-1}}{\sum_h \sum_i BW_{hi0}} \text{ X } 100.0$$

BW_{hit-1} = is the cost weight $(P_{t-1}Q_o)$ for the ith item stratum in the hth PSU at time t-1;

BW_{hi0} = is the base period cost weight (P_oQ_o) for the ith item stratum in the hth PSU;

$R_{hit,t-1}$ = is the price relative for time t for the ith item stratum in the hth PSU.

Indexes for given PSU's, combinations of PSU's, or selected item strata may be estimated by computing the above summation over the corresponding PSU's and item strata.

Introduction to CPI Variance Estimation

The complex sample design of the CPI makes variance estimation by the traditional analytical approach prohibitively costly, if not altogether impossible. Historically, statisticians faced with the same problem with other complex sample designs have developed alternative procedures for variance estimation. The use of independent replications of the sample design or random groups was one of the earlier methods employed. Replication was introduced at the Census Bureau by W.N. Hurwitz, M. Gurney and others. W.E. Deming has consistently advocated replicated sampling.[2/]

The most commonly used replicated sample design calls for two primary sampling units per stratum. This design yields two independent replicated samples with one degree of freedom available for variance estimation. A disadvantage of this design is the lack of a sufficient number of independent replicates to insure sampling stability for estimating variance. The Census Bureau introduced a balanced half-sample replication technique or pseudo-replication to overcome this problem. The technique is also referred to as balanced repeated replication. The mathematics of this approach were developed by P.J. McCarthy and are presented in "Replication: An Approach to the Analysis of Data from Complex Surveys", Vital and Health Statistics, Series 2, Number 14, National Center for Health Statistics.

The following is a simple example of half-sample replication. Suppose we have a stratified sampling procedure with two independent selections made for each of L strata. Let W_h, h=1,2,...,L be the strata weights and y_{hi}, h=1,2,...,L and i=1,2 the observations. Now, a half-sample replicate may be constructed by choosing one member of each pair (y_{11}, y_{12}), (y_{21}, y_{22}), ... (y_{L1}, y_{L2}). There are then 2^L possible half-sample replicates. The population mean is estimated from the entire sample by $\bar{y}_t = \sum_h W_h(y_{h1} + y_{h2})/2$. A half-sample estimate of the population mean is $\bar{y}_{hs} = \sum_h W_h y_{hi}$, i=1 or 2. It has been demonstrated that if k half-samples are independently selected from the 2^L possible half-samples and have means denoted by $\bar{y}_{hs,1}, \bar{y}_{hs,2}, \ldots, \bar{y}_{hs,k}$ then

$$E \sum_{h=1}^{k} (\bar{y}_{hs,i} - \bar{y}_t)^2/k$$

with the expectation taken over the entire set of 2^L half-samples is equal to the variance of $\bar{y}_t$ estimated by the usual analytic method.[3/] As explained by McCarthy, a balanced set of half-sample replicates is obtained by choosing the half-samples in a manner which eliminates the between strata contribution. That is, the half-samples are chosen so that the cross product terms sum to zero.

A balanced half-sample technique will be employed for estimating variances of the CPI. This technique as applied to the CPI is described in the following sections of the paper.

CPI Half-sample Replication

As indicated earlier, 85 geographic areas were selected for pricing for the CPI. The New York Standard Consolidated Area (SCA) was later split into three self-representing areas: New York City, the New York suburbs of New York City, and the New Jersey suburbs of New York City. This yielded a total of 87 areas -- 29 self-representing and 58 non-self-representing.

For publication purposes these areas are divided into four region-size classes. The four regions correspond to the four Census Bureau regions: Northeast, North Central, South, and West. The size classes within each region are: A--self-representing areas; B--other SMSA's with more than 400,000 inhabitants; C--SMSA's with fewer than 400,000 inhabitants; and D--urban areas outside of SMSA's. For publication, the three New York areas are combined. Local area price indexes are published for all self-representing areas. Price indexes are published for the region size class level for the B, C, and D areas.

Prices are collected monthly for all items and indexes are published monthly for the five largest A areas: New York, Los Angeles, Chicago, Philadelphia and Detroit. Prices for food, gasoline and a small number of other items are collected monthly in all other areas. Prices for all other items are collected bi-monthly in these areas. Indexes for the remaining A areas are published bi-monthly with half of the areas on cycle in any given month. Indexes are published monthly for the B, C, and D region size classes. Prices collected monthly are used each month in the estimates of the region size class indexes. Prices collected bi-monthly are used in the region size class indexes only in those months for which the pricing area is on cycle.

As indicated in the section on item sampling, 32 item half-samples were independently selected for pricing. Eight item half-samples were selected for each of the four regions. Each self-representing area (A size class) was randomly assigned two item half-samples for pricing from the eight available for the corresponding region. Each non-self-representing area (B, C, and D size classes) was randomly assigned one of the item samples from the eight available for the corresponding region. The ideal approach would have been to select independently two half-samples for each A area and one half-sample for each B, C, and D area. This would minimize the correlation from pricing the same items across pricing areas. However, workload limitations made a compromise necessary. The 32 independent selections account for the major portion of the possible reduction of the correlation between item samples.

By matching the 87 geographic sample units with the 32 item samples in the above fashion, we have created 16 area/item sample pairs or "strata" for the half-sample replication. This replicate design will yield 2^{16} = 65536 possible half-sample replicates. Balanced half-sample replication requires the independent selection of a subsample of all possible replicates. This sample must have the property that the cross product terms sum to zero, that is it must be orthogonal.

It has been shown that the minimum number of replicates required for orthogonality is equal to the number of pairs plus one plus the value required to make the resulting sum evenly divisible by four. Thus, a minimum of 20 half-sample replicates are required for orthogonality under the CPI design. In an attempt to further reduce the correlation between item samples, the set of 20 half-sample replicates which are the complements of the 20 required for orthogonality were also selected.

Replicate Cost Weights (RBW)

Estimation of the variance for the CPI by the method of half-sample replication requires the construction of price indexes for each half-sample replicate. This necessitates the computation of price relatives for each CPI item strata for each half-sample replicate. It also requires the construction of cost weights for each item strata for each half-sample replicate.

The method of construction of the replicate cost weights (RBW's) for the food, commodities, and services item strata varies by PSU type. For the 26 self-representing PSU's, excluding New York, there will exist two RBW's for each PSU for each item strata. Initially these RBW's will be equal and will also be equal to the total cost weight for the PSU and item strata. Over time the RBW's will vary as each is updated by the price relative estimated from the quotes priced for the corresponding item half-sample.

For New York for each item stratum there will be eight RBW's. There will be two RBW's for each item stratum for each of the three PSU's comprising the New York sample area. In addition, there will be two RBW's for each item stratum for the combined New York area. Again, each RBW will initially be equal to the total cost weight for the New York area. They will vary as they are updated by their corresponding price relatives.

For the twelve region size classes the computation of the RBW's is more complex than for the self-representing PSU's. The number of RBW's required depends on whether the item stratum is priced monthly or bimonthly and on the number of PSU's in the region size class. Each food item stratum and the item stratum for gasoline is priced each month. For each of these item stratum there will be four RBW's for each of the following region size classes: the Northeast's B, C and D size classes; the North Central's B and D size classes; the South's D size class; and the West's B, C and D size classes. For each of these item stratum for the North Central's C size class, there will be eight RBW's. For each of these item stratum, there will be sixteen RBW's for the South's B and C size classes. Initially, each RBW will equal the total cost weight for the item stratum for the corresponding region size class. They will vary as they are updated by the price relatives computed from the corresponding item half-samples.

For each item stratum priced bimonthly, half of the PSU's comprising a region size class will be priced in a given month. Thus, there are two bimonthly pricing cycles: the even numbered months and the odd numbered months. For the Northeast's B, C and D size classes; the North Central's B and D size classes; the South's D size class; and the West's B, C and D size classes there will be two RBW's for each cycle for each bimonthly item stratum. For the North Central's C size class and the South's B and C size classes there will be four RBW's for each cycle for each bimonthly item stratum. For each bimonthly item stratum, there are two total cost weights--one corresponding to each cycle. Initially, each RBW for a region size class will be equal to the corresponding total cost weight for the cycles. Again, the RBW's will vary as they are updated by the price relatives generated from the corresponding item half-samples.

The RBW's for the mortgage interest component are computed as follows. Each sample lending institution is assigned to one of 32 random groups (16 pairs of random groups). Forty half-sample replicates are created for each self-representing PSU and for each of the two bimonthly cycles in the B, C and D region size classes using the bimonthly half-sample replication design matrix. Thus, for each A PSU and for each of the two bimonthly cycles in the B, C and D region size classes there will be forty half-sample replicates -- each requiring a RBW. For the A PSU's the forty RBW's will initially be equal to the total cost weight for each PSU. In the B, C and D size classes, each RBW will initially equal the total cost weight for the corresponding bimonthly cycle. Thereafter, the RBW's will vary as each is updated by the price relatives from the corresponding half-sample replicate.

For rent and property tax each sample housing unit was randomly assigned a replicate pairing code ranging from one to twelve. For each A PSU (self-representing) housing units that have replicate pairing codes 1, 3, 4, 6, 8, 11 are designated as members of the first or A half sample. Housing units with replicate pairing codes 2, 5, 7, 9, 10, 12 are designated as members of the second or B half-sample. Each half-sample is assigned an initial RBW equal to the total cost weight for the PSU. For the B, C and D region size classes, each replicate is assigned an initial RBW equal to the total cost weight for the region size class. Thereafter, all RBW vary as they are updated by the price relatives generated from each half-sample replicate.

Since the universe of FHA house sales is priced for the house price component, no variance estimates are produced for this component. The RBW for a given index are defined to be equal to the total cost weight at any point in time.

The RBW's defined above support estimates of variance of price indexes for each individual item stratum or component at the local level. To support variance estimates of price indexes computed across item strata and index areas, the RBW's must be adapted as follows. The bimonthly half-sample replicate design is used for variance estimates for combinations of item strata and index areas, for example, the total U.S. CPI. Thus, the item strata priced monthly, food and gasoline, must be adapted to the bimonthly design. This is accomplished by summing the corresponding RBW's from the monthly design to the corresponding levels required by the bimonthly design. Estimates for indexes which are combinations of item strata also require the summation of the corresponding RBW's to the combined level.

Computation Formulas for the CPI Variance

Let R_{ijt} be the price relative from time 0 to t for the ith item stratum for the jth half-sample replicate. Let RBW_{ijt-1} be the replicate cost weight for the ith item stratum for the jth half-sample replicate for time t-1. Then, the national all items CPI may be expressed as:

$$I_t = \sum_j \sum_i RBW_{ijt-1} \cdot R_{ijt} \quad \sum_j \sum_i RBW_{ijt-1} \text{ X } 100.0$$

The all items index for the jth half-sample replicate may be expressed as:

$$I_{jt} = \sum_i RBW_{ijt-1} \cdot R_{ijt} \quad \sum_i RBW_{ijt-1} \text{ X } 100.0$$

The price change for the national level for all items between times t and t+k may be expressed as:

$$I_{t+k:t} = I_{t+k}/I_t \text{ X } 100.0$$

Similarly, the price change for all items for the jth half-sample replicate between times t and t+k may be expressed as:

$$I_{jt+k:t} = I_{jt+k}/I_{jt} \text{ X } 100.0$$

The variance of the estimate of price change for the national all items index is expressed as:

$$\text{Var}(I_{t+k:t}) = 1/40 \sum_{j=1}^{40} (I_{jt+k:t} - I_{t+k:t})^2$$

The variance of the estimate of price change for all items for geographic areas below the national level is expressed as:

$$\text{Var}(I_{t+k:t}) = 1/N \sum_{j=1}^{N} (I_{jt+k:t} - I_{t+k:tk})^2$$

where N is the number of replicates. $I_{t+k:t}$ is the price change for the corresponding local area.

In order to compute the variance of the index or of price change for any combination of item strata, it is necessary, first, to compute the index or price change for the combination of item strata. Second, the index or price change must be computed for the half-sample replicates for the combination of item strata. The variances may then be estimated by applying the appropriate equation from above.

Limitations of the Variance Estimates

Balanced half-sample replication has been widely investigated as a method of computing variances for linear estimators. While the basic characteristics of the technique have been established for the linear case, much less is known of the characteristics of the technique for the non-linear case. The technique is seldom used for linear estimators, though, since variances are usually available for the linear case through direct analytical methods. The value of the technique is that it permits relatively easy computation of variance estimates for complex sample designs with non-linear estimators The CPI is, of course. a complex sample design with a non-linear estimator.

Intuitively, the technique seems to mirror all of the complexities of the CPI sample design. However, it is not clear that it accounts for all of the features of the design. For example, the question has been raised as to whether the method accounts for the correlations due to more than one item being priced in the same sample of outlets. This and other questions remain to be investigated.

McCarthy examines some characteristics of the technique for the non-linear case in his paper "Pseudoreplication: Further Evaluation and Application of the Balanced Half-Sample Technique", Vital and Health Statistics, National Center for Health Statistics, Series 2, Number 31. McCarthy's approach is to use the survey data to investigate the variance technique itself. When the CPI variance estimation system becomes operational. it will support similar empirical studies We may then be able to establish some additional characteristics of the technique as applied to the CPI and to non-linear estimators in general..

FOOTNOTES:

[1] A more detailed description of the CPI sample design may be found in "Sample'Design for the Consumer Price Index", by Curtis A. Jacobs in the Proceedings of the 1978 Annual Meeting of ASA.

[2] Deming, W. E., Sample Design in Business Research. New York: John Wiley and Sons, 1960

[3] McCarthy, P. J., "Replication: An Approach to the Analysis of Data from Complex Surveys", Vital Health Statistics, Series 2, Number 14, pp. 16-18, National Center for Health Statistics, Public Health Service, Washington, D.C. 1966.