Reduction of Sample Areas in the Consumer Price Index and Consumer Expenditure Survey Designs

December 2007

Lawrence R. Ernst, William H. Johnson, William E. Larson
Bureau of Labor Statistics, 2 Massachusetts Ave., N.E., Room 1950, Washington, DC 20212-0001
Ernst.Lawrence@bls.gov

Abstract

The Consumer Price Index (CPI) and the Consumer Expenditure Survey (CE) are surveys with multistage designs, revised every 10 years. The first stage CPI and CE samples include a set of areas (PSUs) selected from the set of U.S Core Based Statistical Areas. CE additionally selects a set of PSUs to represent the rest of the nation. After selecting the original sample of PSUs, a reduction was considered for budgetary reasons, a reduction implemented in CE only. In this paper we describe: the details of the reduction process used and alternative approaches; the adjustment of the PSU weights resulting from the reduction, which was complicated by the use of a maximization of overlap procedure in the original selection of the new sample PSUs; and possible improvements to the overlap procedure in the next redesign.

KEYWORDS: PSUs, maximization of overlap, PSU weights

1. Introduction

Every 10 years the Bureau of Labor Statistics has updated the area samples for the CPI (Consumer Price Index) and the CE (Consumer Expenditure Survey). The CPI uses data from CE in order to produce aggregation weights used in constructing higher level indexes. The CE derived weights are used in aggregating across item strata and index areas. For instance, aggregating bananas over all sampled areas in order to produce a U.S. level banana index.

Because of this usage of CE data in the CPI, the desire has been to have the same sample areas for both the CE and CPI. In addition there has been concern about potential bias in the C-CPI-U (chained CPI for urban consumers) if the CE and CPI have different sample areas.

CE covers the entire U.S. population within the 50 states and the District of Columbia. CPI-U covers the portion of the area covered by CE that is in CBSAs (Core Based Statistical Areas). Thus the CPI area sample is normally a subset of the CE area sample.

There are several steps in determining the area sample:

1. A set of certainty areas has to be determined. This is done by means of a population cutoff. The cutoffs used or examined in previous area samples were 1,200,000, 1,500,000, and 1,800,000. As the CPI divides its total sample into 120 units called variance PSUs, we examined using total CBSA population divided by 120, that is 2,142,306, as a cutoff. Although, simulations indicated that with a higher cutoff we should expect lower variances resulting from fewer certainty areas and more non-certainty areas, 2,142,306 was chosen in order not to lose a large number of the current certainty areas.

2. Simulations using components of variance of price change from current samples were used to determine the optimal number of non-certainty PSUs to select in each cluster given the set of certainty areas and the constraint of 120 variance PSUs total. (There are eight clusters, which are the cross product of the four census regions and the two types of areas used in CPI, metropolitan and micropolitan.)

3. Each cluster of non-self representing areas was partitioned into sampling strata with the number of sampling strata equal to the number of PSUs to be selected in the corresponding cluster using a stratification procedure which attempted to minimize a sum of squared differences using several values from the 2000 Census and latitude and longitude. The number of strata in a cluster was the number of PSUs determined to be optimal in step 2.

4. Overlap maximization procedures were run. Since the cost of introducing a new area to the CE/CPI sample is large, there is a desire to have as much overlap as possible, that is to retain as many of the current sample PSUs in the new sample as possible while preserving the unconditional selection probabilities of the PSUs in the new design. The following procedure was used to produce the adjusted probabilities of selection:

The first step was to determine what is meant by an overlap PSU. Given the considerable changes in definitions of the PSUs it is possible that part of a PSU might currently be in the CPI sample but not other parts. The preliminary definition was that 30% of the counties or 30% of the 2000 population of a PSU currently be covered by the CPI sample. This was complicated by the fact that counties are composed of Minor Civil Divisions (MCDs) in the Northeast region. Current CPI PSUs in the Northeast are defined at the MCD level, while the new PSUs are defined at the county level. It was decided that a county composed of MCDs was overlap if at least 5% of its 2000 population was covered by the current CPI sample. A PSU composed of MCDs is considered overlap as long

as 30% of the counties are overlap and at least one of those counties is at least 30% covered by the current CPI sample based on number of MCDs.

The overlap maximization procedure, which is a modification of the procedure used in the previous redesign, attempts to increase the likelihood of selecting PSUs which are overlap. Some changes in the program had to be made due to the massive redefinition of PSUs. The procedure used operates at the level of the intersection of a new stratum with a stratum for the 1998 CPI Revision PSU sample. Due to redefinitions, there are many cases where only part of a new PSU lies within one of these intersections. Thus the new PSUs were broken in pieces for the purpose of overlap maximization and then the pieces were added together to give the total new probability of selection of a PSU. See Section 4 of this paper for a further description of this procedure.

5. A controlled selection program was run to construct a complete set of patterns in each Census region and then select a pattern in each region. Among the controls were the requirements that:

Each state had to receive the number of PSUs it was expected to have based on population (rounded up or down to one of the two nearest integers).

In each region the number of overlap PSUs had to equal the expected number of overlap PSUs based upon the adjusted probabilities of the overlap PSUs (rounded up or down).

In each cluster of sampling strata, the number of PSUs selected had to equal the expected number of PSUs selected from the cluster (rounded up or down.)

The unconditional probability of any new PSU being selected had to remain unchanged by the controlled selection procedure.

The resulting selected sample of 86 areas, which will be referred to as the 86 PSUs design, was to be used for the CPI, with the CE using these 86 areas and 16 additional areas that are outside CBSAs. CE data for the new area sample was collected by Census starting in 2005. The CPI was originally supposed to start introducing the new area sample in 2008, which would replace the current area sample, an 87 PSUs design.

CPI's initial request for funding for introducing new areas and a continuous rotation of the housing sample was denied. It was decided that the cost of the CPI revision should be brought down. CPI does its own data collection so it was felt that the greatest savings would be realized by reducing the number of areas.

A description of the sample cut procedure is presented in Section 2. The remaining sections focus on the overlap procedure, principally because it affects the changes in the PSU weights due to the sample cuts. In Section 3 a linear programming overlap procedure is described, which has some advantages over the currently used procedure. In Section 4 a slightly modified version of the overlap procedure actually used in the selection of the PSUs in the 86 PSUs design is presented. The changes, which are mainly changes in presentation, were principally done to simplify comparisons between the procedure actually used in selecting the PSUs for the 86 PSUs design and the linear programming procedure. The original description of the overlap procedure actually used is presented in Johnson, Shoemaker, and Rhee (2002). It is a generalization of a procedure due to Perkins (1970). Finally, in Section 5, the calculations of the PSU weights after the sample cut are described. In particular, these weights are affected by the use of an overlap procedure.

2. Description of Sample Cut Procedure

The CPI program decided to reduce the number of sample areas in the new area sample from 86 to 75 while keeping the same overall quote and housing units sample sizes by increasing within PSU sample sizes. It was believed that eliminating the overhead and startup costs and not having to maintain staff in these 11 areas would save approximately $1,000,000 per year.

As the most expensive areas to collect data in the new design are the areas which are not in the current CPI and therefore are areas for which we do not already have data collection staff and offices in place, the only areas considered for being cut were newly selected areas.

In the CPI area sample, PSUs can be A-size (self representing), X-size (mostly metropolitan CBSAs with a few exceptions), and Y-size (mostly micropolitan CBSAs). Given the increasing population share of the Y-size areas and their relatively small samples, only the X-size areas were considered for being dropped. Note also that in the CPI's staff documentation of the sample cut process, several of the A PSUs were changed to X PSUs in conjunction with the sample cut since it was intended that indexes would no longer be published for these areas. However, since these PSUs remain certainty in the 75 PSUs design, we consider them in this paper to remain A PSUs after the sample cut.

In the Northeast region, there was one newly selected PSU and it was cut.

In the Midwest region, all three newly selected PSUs were dropped.

In the South region, there were 10 newly selected areas. Three of these areas were dropped, with these three areas selected from the 10 with equal probability. This is the only region in which some of the newly selected PSUs were not cut.

In the West region, all four newly selected PSUs were dropped.

Thus a new sample of 75 PSUs for the CPI was determined. The PSU weights of the X sample PSU remaining in the 75 PSUs design were increased to reflect that these PSUs are only in sample for the 75 PSUs design if they are in sample for the 86 PSUs design and they are retained in the sample cut. The details of the weight adjustments are presented in Section 5.

Due to budget problems, CE needed to save about $1.2 million and in 2006 implemented this reduced sample by having the Census Bureau, which collects the data for CE, drop data collection in the 11 areas CPI decided to cut and to adjust the PSU weights in the same way CPI intended.

CPI ultimately decided to return to the originally selected 86 PSUs sample, deciding that it could make up the estimated $1,000,000 savings in other ways. As of this time, Congress has not yet approved CPI revision funding to implement a new area sample.

3. Overlap Procedure of Ernst (1986)

As mentioned in the Introduction, the overlap procedure used in the selection of the CPI PSUs in the 86 PSUs designs and earlier designs is a generalization of a procedure due to Perkins (1970).

An alternative is an overlap procedure due to Ernst (1986), which has at least four major advantages over Perkins' procedure, namely:

A. It generally yields a higher expected overlap than Perkins' procedure and never yields a lower expected overlap.

B. It is easy to modify this procedure to handle the case when the PSUs are redefined in the new design.

C. The procedure does not require that either the initial design or the new design be one PSU per stratum. In this paper, initial design and new design mostly refer to the 87 PSUs and 86 PSUs CPI designs, respectively.

D. The procedure can handle many different types of rules for what is considered an overlap, or what is considered a partial overlap. (See the last paragraph of this section for a discussion of partial overlap.)

In addition, the Ernst (1986) procedure shares with Perkins' procedure the desirable property that it is not required that the PSUs in the initial sample be selected independently from stratum to stratum. The reason that this property is mentioned is that there are some overlap procedures, such as the procedure of Causey, Cox, Ernst (1985), that require this independence assumption but do not preserve this independence from stratum in the new design and hence cannot be used in two consecutive redesigns. Therefore, these overlap procedures are not described here. This issue is discussed in more detail in Ernst (1986).

We proceed to describe the procedure of Ernst (1986). Since the CPI is a one PSU per stratum design, we restrict the description to this case in order to simplify it, although this restriction is not necessary and is not used in Ernst (1986). Also, since in the CPI the PSUs definitions differ in the 87 and 86 PSUs designs, we do not restrict our description to the case when the PSU definitions are identical in the old and the new designs.

Note first that each stratum *S* in the new design represents a separate problem, with $S_k, k = 1,..., N$, denoting the PSUs in *S* and π_k denoting the probability of selection of S_k in the new sample. Let $I_1,..., I_r$ be the strata in the initial design having at least one PSU that intersects a PSU in *S*. The general idea of the procedure is to select one of the I_i and then condition the selection of the new sample PSU from *S* on which PSU from I_i was chosen in the initial sample. Furthermore, in general, one specific I_i is not chosen with certainty, but instead a probability y_i is assigned to each of the I_i, with the y_i, $i = 1,..., r,$ being a set of variables in the optimization process. For each $i = 1,...r,$ let I_{ij}, $j = 1,..., u_i,$ denote the PSUs in I_i; let x_{ijk} denote the joint probability that I_i is the selected initial stratum, I_{ij} is the PSU from this stratum selected in the initial sample, and S_k is the new PSU selected from *S*. Let p_{ij} be the probability of selection of I_{ij} in the initial sample.

The x_{ijk}'s are variables in the linear programming problem to be described and the only such variables besides the y_i's. After an optimal set of values is obtained by solving the linear programming problem, the desired probabilities, which are the probabilities of selection of each of the S_k's conditioned on the initial sample PSUs in $I_1,..., I_r$, can be expressed in terms of the optimal y_i's and x_{ijk}'s and the known p_{ij}'s as follows.

$$\Pr(S_k \mid I_i, I_{ij}) = \frac{\Pr(I_i, I_{ij}, S_k)}{\Pr(I_i, I_{ij})} = \frac{x_{ijk}}{y_i p_{ij}},$$

from which it follows that

$$\Pr(S_k \mid I_{1j_1},....,I_{rj_r}) = \sum_{i=1}^{r} \frac{x_{ij_i k}}{p_{ij_i}}$$

where I_{ij_i} $i = 1,..., r,$ is the PSU in stratum I_i that was

in the initial sample.

All that remains is to state the linear programming problem that yields the optimal set of x_{ijk}'s and y_i's. The constraints will be presented first and then the objective function. Since new unconditional probabilities must be preserved, the probabilities of all the three stage events with S_k as the new sample PSU in *S* must sum to π_k, that is

$$\sum_{i=1}^{r}\sum_{j=1}^{u_i} x_{ijk} = \pi_k,\ k = 1,...,N$$

Similarly, since $\Pr(I_i, I_{ij}) = p_{ij} y_i$, we also have that

$$\sum_{k=1}^{N} x_{ijk} = p_{ij} y_i,\ \ i = 1,...,r,\ \ j = 1,...,u_i$$

The final constraint is

$$\sum_{i=1}^{r} y_i = 1$$

which arises from the fact that exactly one initial stratum is chosen to provide information that is input in the process of selecting a new sample PSU from *S*.

As for the objective function, it will be of the form

$$\sum_{i=1}^{r}\sum_{j=1}^{u_i}\sum_{k=1}^{N} c_{ijk} x_{ijk}$$

where c_{ijk} for each i, j, k is a constant to be determined.

One possibility is for c_{ijk} to be either 0 or 1 depending on the magnitude of the intersection of I_{ij} and S_k; that is outcome i, j, k leads to overlap and hence $c_{ijk} = 1$ if, for example, a certain percentage of S_k is covered by I_{ij} based on either the new design population or the number of subdivisions of the PSU, such as counties, while otherwise $c_{ijk} = 0$. This approach is based on the requirement of a single initial sample PSU covering a certain percentage of the new sample PSU rather than an alternative of all the initial sample PSUs intersecting S_k combined covering this percentage.

A second possibility is for each c_{ijk} to be determined by considering PSU I_{ij} together with a set of initial PSUs with one initial PSU from each initial stratum other than I_i that intersects *S*. This yields a set of PSUs one from each stratum $I_1,...,I_r$. For each such set the product of the initial selection probabilities of each PSU in this set except for PSU I_{ij} is calculated and the product summed over all such sets for which the set covers the required percentage of S_k to obtain c_{ijk}. This is the desired goal of the approach taken in the selection of sample PSUs for the 86 PSUs CPI design, since for this selection a PSU (except for a PSU composed of MCDs) is considered to be overlap if 30 % of the 2000 population of the PSU or 30% of the counties in the PSU are covered by the 87 PSUs sample, regardless of the number of PSUs from the 87 PSUs design needed to meet this 30% figure. However, for the procedure actually implemented, as described in Section 4, this overlap goal may actually not be met since that procedure attempts to retain the intersection of an initial sample PSU and a new PSU regardless of how little of the new population of the new PSU is in the intersection.

A third approach is to have the value of c_{ijk} be the proportion of S_k that is in the intersection of I_{ij} and S_k based on 2000 population. This approach is illustrated in Ernst (1986). This approach, unlike the other two approaches allows for an outcome i, j, k to be considered a partial overlap, that is $0 < c_{ijk} < 1$, when an initial sample PSU partially intersects a new sample PSU.

4. Perkins' Method and its Generalization

The method of Perkins (1970) has some of the same characteristics as that of Ernst (1986), at least for one PSU per stratum designs, which are the only type considered in this paper. (Perkins never generalized his procedure to other designs, while the Ernst (1986) procedure is not restricted to such designs.) In addition, in Perkins' procedure, unlike the Ernst procedure, it is assumed that that the PSU definitions are the same in the initial and new designs. Since for CPI the PSUs in the 87 PSUs design and the 86 PSUs design are defined differently, the selection of PSUs for the 86 procedure was done using a generalization of Perkins' procedure due to Johnson, Shoemaker, and Rhee (2002), which will be referred to as JSR. This procedure allows for different PSU definitions in the initial and new designs. However, although in the CPI application a new PSU is generally considered to be overlap if it is 30% covered by the initial sample, as explained in the Introduction, JSR actually attempts to select any new PSU that has any intersection at all with an initial sample PSU.

Although the procedure of Ernst (1986) has several advantages as we have mentioned in Section 3, we proceed to describe JSR since this was the procedure used in selecting the 86 PSUs design and hence is needed in calculating the probability of a PSU in the 86 PSUs design being considered overlap. As mentioned in the Introduction, our description of the JSR procedure will be somewhat different than given

in that paper.

For $i = 1,...,r,\ \ j = 1,...,u_i,\ k = 1,,,,,N,$ let π_{ij} be the proportion of the 2000 population of *S* that is in I_{ij} and let $\pi_{k|ij}$ be the proportion of the 2000 population of $I_{ij} \cap S$ that is in S_k. y_i has the same meaning as in Section 3, which is the probability of initial stratum I_i being selected to provide information on the initial sample. However, for JSR y_i is not obtained by solving a linear programming problem but instead is simply the proportion of the 2000 population of *S* that is covered by I_i.

Also, instead of calculating the variables x_{ijk} using linear programming, we calculate x_{ijtk} for $i = 1,...,r,\ \ j,t = 1,...,u_i,\ k = 1,,,,,N,$ where x_{ijtk} is the joint probability of the following four events: the piece of the new PSU in *S* is selected from I_i; I_{ij} is the initial sample PSU selected from I_i; the new PSU in *S* is selected from a piece of a PSU in *S* that resides in I_{it}; and the piece of the new PSU that is selected from I_{it} resides in S_k, that is S_k is the new sample PSU selected from *S*. x_{ijkt} is obtained by letting

$$x_{ijjk} = y_i p_{ij} \min\{\pi_{ij}/(y_i p_{ij}),1\}\pi_{k|ij}$$

$$x_{ijtk} = \frac{y_i p_{ij}(1 - \min(\pi_{ij}/(y_i p_{ij}),1\})\pi_{k|it}}{\sum_{l=1}^{u_i}(\max\{\pi_{il} - y_i p_{il},0\})} \times \max\{\pi_{it} - y_i p_{it},0\} \quad \text{if } t \neq j$$

where $\pi_{k|it}$ is the proportion of $S \cap I_{it}$, based on the 2000 population, that resides in S_k. It then follows that

$$\Pr(S_k \mid I_i, I_{ij}) = \frac{\sum_{t=1}^{u_i} x_{ijtk}}{y_i p_{ij}}$$

and hence

$$\Pr(S_k \mid I_{1j_1},...,I_{rj_r}) = \sum_{i=1}^{r} \frac{\sum_{t=1}^{u_i} x_{ijtk}}{p_{ij_i}}$$

where I_{ij_i} $i = 1,...,r$, is the PSU in stratum I_i that was in the initial sample.

5. Calculation of PSU Weights After Sample Cut

In order to obtain PSU weights after the sample cut that reflect the probability of selection, the key is that this weight, which is a random variable *w*, must satisfy $E(w) = 1$ for each of the 75 retained PSUs and, in addition, that $w = 0$ for any PSU not among the 75 PSUs. See Ernst (1989) for a discussion of this issue. Here the expectation is over all possible 87 PSUs design sample PSUs, 86 PSUs design sample PSUs selected using the overlap maximization procedure, and 75 PSUs design sample PSUs selected using the sample cut procedure.

We assume that the weight before the sample cut satisfies the conditions that $E(w) = 1$ for each of the PSUs in sample for the 86 PSUs design and $w = 0$ if the PSU is not in sample for the 86 PSUs design. We only consider sample PSUs that are X PSUs in the 86 PSUs design, since these are the only PSUs that are eligible to be cut in the 75 PSUs design. We also assume that the sample PSUs in the 87 PSU design were selected independently from stratum to stratum. This is not actually true since the sample PSUs in the 87 PSUs design were selected using an overlap maximization procedure and, as shown in Ernst (1986), the use of most overlap maximization procedures, including Perkins' procedure, destroys the stratum to stratum independence. See Case 2 below for further discussion of this issue.

We proceed to describe the weighting procedure for the 31 PSUs that are X sample PSUs in both the 86 PSUs design and the 75 PSUs design by considering the following cases. (In general more complex cases are possible, but in this work only the cases described below occurred.)

Case 1. South region.

Case 1.A. X PSUs in 86 PSUs design that overlap PSUs in 87 PSU design.

There are 8 such PSUs. The PSU weight for each of these PSUs after the sample cut is the same as it is before the sample cut since each of these PSUs are retained in the 75 PSU design with certainty.

Case 1.B. All other X PSUs.

There are 10 sample X PSUs in the South region in the 86 PSUs design that are not considered as continuing from the 87 PSUs design. Of the 10, 7 were selected with equal probability to remain in sample in the 75 PSUs design. For each of these 7 PSUs, the PSU weight before the cut is multiplied by 10/7 to obtain the base PSU weight after the cut. This weight is adjusted by a benchmark factor, which forces the sum of the weighted PSU population in the 7 retained sample PSUs in the South region to agree with

the 2000 census population for the 10 strata from which these 7 PSUs were selected. Similar benchmark factor adjustment would be done in the other three regions.

Case 2. Three other regions.

For each of the three subcases of Case 2 described below, we present a formula for the probability that an X PSU was in sample for both the 87 PSUs design and the 86 PSUs design, since in these three regions an X PSU is in the 75 PSU sample if and only if these two events both occur. The reciprocal of this joint probability is the base PSU sample weight after the sample cut. Now if the selection of a PSU in the 87 PSUs design and the selection of a PSU in the 86 PSUs design were independent events, then this joint probability would be the product of the probability of being in the 87 PSUs design and the probability of being in the 86 PSUs design. However, this independence does not hold because, as explained earlier, a procedure was used in selecting the 86 PSUs design sample that maximized overlap with the 87 PSUs design sample and in general the use of such an overlap procedure destroys such independence. For the three sub-cases we present expressions for the joint probability that correctly take into account the use of the overlap maximization procedure in the selection of the PSUs in the 86 PSUs design. However, to avoid undue complexities, the calculation makes the incorrect assumption that the sample PSUs in the 87 PSUs design were selected independently from stratum to stratum, when in fact these PSUs were also selected using an overlap procedure. In addition, we ignored the fact that a controlled selection procedure, as explained in JSR and summarized in Step 5 of Section 1 of this paper, was used in selecting the PSUs in both the 87 and 86 PSUs designs and that controlled selection generally destroys the independence of sampling from stratum to stratum. As a practical matter we had no choice, since as far as we know the data to completely calculate the joint probabilities correctly no longer exists. We believe that taking into account the use of the overlap procedure in selecting the sample PSUs for the 86 PSUs design, as we have done, is the most important step that we could have taken in obtaining the correct joint probabilities.

We now proceed to present these three sub-cases:

Case 2.A. An entire sample PSU S_k in the 86 PSUs design is contained in a single sample PSU I_{ij} in the 87 PSUs design. There are 11 S_k in these regions satisfying this condition.

For each such S_k, let p_{ij} be the probability of selection of the corresponding I_{ij} in the 87 PSUs design and let π_{ij} be probability that the PSU selected from S in the 86 PSUs design is selected from a piece of I_{ij}. Let y_i be as described in Section 4. Then the probability that S_k is an overlap PSU, which is the probability that I_{ij} is in sample for the 87 PSUs design and S_k is in sample for the 86 PSUs design is

$$y_i p_{ij} \min\{\pi_{ij} /(y_i p_{ij}), 1\}\pi_{k|ij} \qquad (5.1)$$

The reason that this is so is that $y_i p_{ij}$ is for stratum S the joint probability that I_i is the corresponding selected initial stratum and I_{ij} is the initial sample PSU in I_i for the 87 PSUs design, since these two events are independent. $\min\{\pi_{ij} / y_i p_{ij}, 1\}$ is the probability that the new PSU in S will be selected from a piece of I_{ij} in the 86 PSUs design given the first two events and $\pi_{k|ij}$ is the probability that the piece of I_{ij} selected is S_k.

Case 2.B. Two of the X PSUs, S_k, in the 86 PSUs design have the following properties. Each of these PSUs was included in a single sample PSU I_{ij} in the 87 PSUs design sample except that S_k contains counties that were not included in an urban area in the 87 PSUs design and hence were not in I_{ij} and not eligible for the CPI sample.

Then (5.1) can be applied in this case as in Case 2.A if we artificially consider the non-urban counties that are part of S_k in the 86 PSUs design to be part of I_{ij} in the 87 PSUs design but with an urban population of 0. That is, under these conditions, S_k in the 86 PSUs design is contained in a single sample PSU I_{ij} in the 87 PSUs design and (5.1) applies.

Case 2.C. There are three PSUs in the 86 PSUs design that constitute this case. Each of the PSUs in this design consists of two urban counties, county 1 and county 2, which in the 87 PSUs design were in different urban strata.

Let I_1, I_2 be the 87 PSUs design strata containing county 1 and county 2, respectively; let I_{11}, I_{21} be the PSUs in this design containing county 1 and county 2, respectively; and let S be the 86 PSUs design stratum containing counties 1 and 2. Let p_{11}, p_{21} be the probability of selection of I_{11}, I_{21} in the 87 PSUs design and let π_{11}, π_{21} be the proportion of the 2000 population of S in I_{11}, I_{21}, respectively. Let

y_i, $i = 1,2$, be the proportion of the 2000 population of S covered by I_i. Then the probability that county 1 and/or county 2 is in the 87 PSUs design sample and the PSU S_k consisting of counties 1 and 2 is in the 86 PSUs design sample is

$$y_1 p_{11} \min\{\pi_{11}/(y_1 p_{11}), 1\}\pi_{k|11} + y_2 p_{21} \min\{\pi_{21}/(y_2 p_{21}), 1\}\pi_{k|21}$$

References

Causey, B. D., Cox, L. H., and Ernst, L. R. (1985). Applications of Transportation Theory to Statistical Problems. Journal of the American Statistical Association, 80, 903-909.

Ernst, L. R. (1986). Maximizing the Overlap Between Surveys When Information Is Incomplete. European Journal of Operational Research, 27, 192-200.

_____ (1989). Weighting Issues for Longitudinal House and Family Estimates. Panel Surveys, 139-159. New York, John Wiley.

Johnson, W. H., Shoemaker, O. W., and Rhee, Y. W. (2002). Redesigning the Consumer Price Index Area Sample. Proceedings of the Section on Government Statistics, American Statistical Association, 1671-1676.

Perkins, W. M. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Strata. Memorandum to Joseph Waksberg, U.S. Bureau of the Census.

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