SELECTING THE EMPLOYMENT COST INDEX SURVEY SAMPLE AS A SUBSAMPLE OF THE NATIONAL COMPENSATION SURVEY

Lawrence R. Ernst and Chester H. Ponikowski, Bureau of Labor Statistics
Chester H. Ponikowski, BLS, 2 Massachusetts Ave. N.E., Rm 3160, Washington, DC 20212

Key Words: ECI, NCS, BLS, Compensation Surveys, Integration

1. Introduction

The National Compensation Survey (NCS) is a new statistical program that will both replace the existing Occupational Compensation Survey (OCS) program and integrate it with the Employment Cost Index (ECI) and the Employee Benefit Survey (EBS), creating one comprehensive survey program. The OCS program publishes locality and national occupational wage data used by the President's Pay Agent and private sector compensation specialists, among others. The ECI publishes national indexes which track quarterly and annual changes in employers' labor costs and also cost level information annually on the cost per hour worked of each component of compensation. The EBS publishes annually incidence and detailed provisions of selected employee benefit plans.

Although the ECI and EBS sampling have been integrated for several years, the OCS and ECI/EBS have been independent samples, collected separately by regional field staff. These survey programs are being combined because of a desire to lessen the respondent burden and to maximize the use of limited resources. Similar to the OCS program, the NCS produces estimates of occupational wages for Locality Pay and constructs national estimates from a probability selection of establishments stratified geographically and by industrial activity. The NCS also will maintain the current products of the EBS and ECI surveys.

One of the primary goals of the integration of these Bureau of Labor Statistics (BLS) surveys is to select future establishment sample for the Employment Cost Index (ECI) survey as a subsample of the larger National Compensation Survey (NCS). The current NCS sample of establishments was selected with probability proportional to size (pps), with total employment being the measure of size, from a frame covering establishments with 50 or more employees. The current NCS sample generally consists of two independent samples, as explained in Section 3, with some establishments selected twice. The conversion of the ECI to a subsample of the NCS will be accomplished by selecting and initiating five subsamples of the current NCS sample over three years.

In designing the subsampling plan for the ECI we came across several issues that may be of interest to survey practitioners who may have a need to select subsamples of a larger establishment sample. Under our plan the subsampling probabilities are determined with the goal that all establishments in an industry stratum in the universe with the same frame employment have the same overall chance of being in the ECI sample, regardless of the geographic PSU in which they are located. Even though the NCS sample is actually two independent samples with some establishments selected twice, the subsampling plan insures that no establishment will be selected more than once in ECI. In addition the plan handles the additional complexities arising from the fact that the current NCS sample consists only of establishments with frame employment of at least 50, while the ECI has no minimum employment restrictions, necessitating supplemental ECI samples of the smaller establishments.

This paper describes the sample design for the NCS (Section 2), our plans for subsampling the NCS sample (Section 3), alternative approaches to sample allocation and selection (Section 4), and selection and weighting of subsamples of the ECI samples (Section 5).

2. Sample Design for the NCS

The NCS sample is selected using a 3-stage stratified design with probability proportional to employment sampling at each stage. The first stage of sample selection is a probability sample of areas, the second stage is a probability sample of establishments within sampled areas, and the third stage of sample selection is a probability sample of occupations within sampled areas and establishments.

The selection of sample areas is done by first dividing the entire area of the United States, consisting of counties and independent cities, into primary sampling units (PSUs). In most States, a PSU consists of a county or a number of contiguous counties. Metropolitan areas, as defined by Office of Management and Budget (OMB), are used as a basis for forming PSUs. Outside of metropolitan areas, each county defines a PSU.

The PSUs with similar average wages as measured by Unemployment Insurance reports wage are grouped

into strata within each of the 9 BLS economic regions. Then one PSU is selected from each stratum with the probability of selection proportional to the employment of the PSU. There are 33 PSUs that are self-representing, and these include the 18 Consolidated Metropolitan Statistical Areas (CMSAs) and 15 largest Metropolitan Statistical Areas (MSAs). The remaining strata are formed by combining PSUs that are MSAs and have similar average annual pay into 45 MSA strata, and PSUs that are non-MSAs and have similar average annual pay into 73 non-MSA strata. The PSUs selected with probability proportionate to PSU employment from these strata are non-self-representing because each one chosen represents the entire stratum. In addition the NCS design was supplemented with 3 PSUs to meet requirements of Federal Employee Pay Comparability Act of 1990, commonly known as Locality Pay. Since these 3 additional PSUs were not part of the original sample design they were not used in selecting the ECI sample.

The sample of establishments is drawn by first stratifying the sampling frame for each PSU by industry and ownership (private, state government and local government). The number of sample establishments allocated to each stratum is approximately proportional to the stratum employment. Each sampled establishment is selected within a stratum with a probability proportional to its employment. The set of industry strata were chosen because of a desire to produce estimates of major industry divisions as well as selected individual industries that are traditionally produced for the ECI and locality products.

After the sample of establishments is drawn, occupations are selected in each sampled establishment. The number of occupations selected in an establishment depends on the total number of employees in the establishment. Anywhere from 4 to 20 occupations are selected within sampled establishments. The probability of an occupation being selected is proportionate to its employment within the establishment. For more detailed description of the NCS sample design, refer to the BLS Handbook of Methods (Bulletin 2490, April 1997) and Black et al (1997).

3. Plans for Subsampling the NCS Sample

For both the NCS and the ECI, the sampling frames are partitioned into a set of industry strata, with the sampling independent from stratum to stratum. Consequently, the ECI sampling process below actually applies to the separate selection of the sample in each industry stratum, although generally we will not be specifically stating this in the discussion. For example, when we speak of "the ECI sample" we really mean the ECI sample in a single industry stratum.

The ECI selection process roughly proceeds as follows. The ECI sample that will enter sample over the next three years will consist of five panels. First a single sample, denoted as the S sample, will be selected. The establishments in this sample with assigned employment ≥ 50 will be selected as a subsample of the NCS sample, which in turn was selected from a UDB frame denoted as the D_0 frame. The remainder of the S sample will be selected directly from the most recent available UDB frame, denoted as the D_1 frame, at the time that the S sample is selected; that is the D_1 frame is more recent than the D_0 frame. Then five subsamples of the S sample, denoted $S_1,...,S_5$, will be selected. Those establishments that are selected into S from the NCS sample with certainty are in each of the $S_1,...,S_5$ samples. The remaining units in S are assigned to exactly one of these samples. These samples are not the final ECI samples, which will be denoted $S_1',...,S_5'$. Now $S_1' = S_1$ and for all k, S_k' will include all NCS sample units in S_k. However, the units in S_k for $k \geq 2$ selected from D_1 will generally not be in S_k'. Instead, the units with assigned employment <50 in S_k' will be selected from an updated frame, D_k. For each sample PSU j the number of establishments, n_{jk}, with assigned employment <50 in S_k' selected from PSU j is the same as the number of such units in S_k. Thus, the selection of units from D_1 in S_k for $k \geq 2$ is only done for the purpose of determining the allocation of the <50 establishments among the sample PSUs in $S_2',...,S_5'$, not for actually selecting the sample establishments.

There are several additional frames involved in selecting the ECI sample. Let G_k, L_k, $k = 0,...,5$, denote the subframes of D_k consisting of establishments with assigned employment ≥ 50 and <50, respectively. Next let $L_k' = (L_k \sim G_0) \cup (G_k \cap L_0)$, and $C_k = G_0 \cup L_k'$, $k = 1....,5$, that is L_k' is a subframe of D_k that can be considered a modification of L_k obtained by removing from L_k those establishments that are also in G_0, and would otherwise be listed twice in C_k, and adding in those establishments in $G_k \cap L_0$, that is units that otherwise would not be listed in C_k since they are in neither G_0 or L_k. C_k is the universe for selecting the S_k' sample, with the assigned employment in its G_0 component taken from the D_0 frame and its assigned employment from the L_k' component arising taken

from the D_k frame. Finally, abbreviate $G = G_0$, $L = L_1'$, and $C = C_1$.

We proceed to discuss in detail the sampling and weighting for the selection of the S sample from the C frame, which involves three stages of sampling. The first stage is the selection of the sample PSUs. Let F denote the set of sample PSUs and for any PSU j, let p_j denote the probability that $j \in F$. Let C' denote the set of establishments in C in sample PSUs. C' is the frame for the second stage of sampling. The weight, w_{Fij}, associated with the first stage of selection is given by $w_{Fij} = 1/p_j$ if $ij \in C'$, $w_{Fij} = 0$ if $ij \in C \sim C'$.

The second stage is the selection of the NCS sample units from establishments in $C' \cap G$. There is an extra complexity here because for most of the sample PSUs, the set of NCS establishments was selected as two independent samples for the following reason. The NCS data collection process for the PSUs for which data collection began early took longer than originally anticipated. As a result, in order to meet scheduled completion dates, a smaller sample than originally intended was selected for the remaining sample PSUs. This reduced size sample, which in most PSUs consisted of 1/2 the original intended number of establishments is known as the A sample. Later it was decided to restore the original sample size in those PSUs where it had been cut. This was accomplished by selecting a second sample, known as the B sample, in each such PSU, with the B sample selected independently of the A sample. As a result, the same establishment may be selected twice for NCS, once in the A sample and once in the B sample. For each sample PSU, both the A and the B samples are systematic pps samples of the G frame, with assigned employment the measure of size. The NCS sample is then the set of all establishments in $A \cup B$. We let C'' be the subset of C' consisting of all establishments in C' that are either in the NCS sample or the L frame. C'' is the frame used in the third stage of sampling.

For each $ij \in C'$, where ij denotes establishment i in PSU j, we associate a weight, w_{Nij}, arising from the selection of the NCS sample from the sample PSUs and hence the selection of the units that are in C''. That is, w_{Nij} is the weight associated with the second stage of sampling. Let $w_{Nij} = 1$ if $ij \in C' \cap L$, since all these units are in C'' regardless of the NCS sample. As for units ij in $C' \cap G$, if ij is in the A sample let w_{Aij} denote the reciprocal of the probability of selecting this unit in the sample A given that $j \in F$; if ij is not in the A sample let $w_{Aij} = 0$. w_{Bij} is defined analogously. Then let $w_{Nij} = \alpha_A w_{Aij} + \alpha_B w_{Bij}$, where α_A, α_B are factors which satisfy

$$\alpha_A + \alpha_B = 1, \tag{1}$$

and

$$\alpha_B / \alpha_A = n_B / n_A, \tag{2}$$

where n_A, n_B are the number of sample establishments in samples A and B, respectively. Requirement (1) is needed to obtain weights that yield unbiased estimates, as will be shown later, while (2) is used to minimize variances for estimates for the combined A and B samples under the assumption that the ratio of the variances for estimates obtained from these two samples are inversely proportional to the ratio of their sample sizes. In some sample PSUs the A sample was not reduced from the originally intended sample size, in which case there is no B sample; hence $\alpha_B = 0$ and $w_{Bij} = 0$ for all establishments in these PSUs in $C' \cap G$. Also let $w_{Nij} = 0$ if $ij \in C \sim C'$, since such units have no chance of selection in C''. The assignment to each establishment of a single weight, w_{Nij}, arising from the selection of the NCS sample from sample PSUs, even if the establishment was selected in both the A and B samples, is the key to insuring that no establishment is selected more than once in ECI.

Let $\Omega = \{w_{Nij} : ij \in C\}$. To select the S sample establishments from C'', that is the final stage of selection, first sort the establishments in C'' by PSU and within PSU by ECI pseudo standard industry code (PSIC), and then assigned employment, denoted T_{ij}. The PSICs partition each industry stratum into a set of finer industries. Then, using the measure of size $w_{Nij} T_{ij} / p_j$, simply select a systematic pps sample of units in C''. For $ij \in S$, let $w_{\Omega ij}$ be the reciprocal of the probability that $ij \in S$ conditioned on Ω, while $w_{\Omega ij} = 0$ for $ij \in C \sim S$. Thus $w_{\Omega ij}$ is the weight associated with the third stage of selection.

The overall weight for each ij in C, denoted w_{Sij}, reflecting the three stages of selection just described is simply the product of the weights for each stage of selection, that is

$$w_{Sij} = w_{Fij} w_{Nij} w_{\Omega ij}. \tag{3}$$

We proceed to demonstrate that the weights w_{Sij} yield unbiased estimates of totals and are as close to

being inversely proportional to T_{ij} as possible given the constraint that $S \cap G$ is a subsample of the NCS sample. In fact it is the desire to have w_{Sij} be inversely proportional to T_{ij} that motivated the use of the measure of size $w_{Nij}T_{ij} / p_j$.

We first show that

$$E(w_{Sij}) = 1, \ ij \in C, \tag{4}$$

which, by Ernst(1989), is sufficient to prove that w_{Sij} yield unbiased estimates of totals. To establish (4), we observe that by the definitions of $w_{Aij}, w_{Bij}, w_{Nij}$,

$$E(w_{Aij} | j \in F) = 1, \ E(w_{Bij} | j \in F) = 1 \text{ if } a_B \neq 0,$$

which together with (1) establish that

$$E(w_{Nij} | j \in F) = 1. \tag{5}$$

Furthermore by the definitions of $w_{Fij}, w_{\Omega ij}$,

$$E(w_{Fij}) = 1, \ E(w_{\Omega ij} | w_{Nij}) = 1 \text{ if } w_{Nij} \neq 0, \tag{6}$$

We combine (3), (5) and (6) to conclude

$$E(w_{Sij}) = E[w_{Fij} E(w_{Nij} | w_{Fij}) E(w_{\Omega ij} | w_{Nij})] = 1.$$

To show that the weights w_{Sij} are as close to being inversely proportional to T_{ij} as possible, we proceed as follows. Let I_Ω be the final sampling interval used in selecting the S sample, where the notation for I_Ω is chosen to indicate the dependence of the sampling interval on Ω. Then for establishment ij in the S sample, $w_{\Omega ij} = p_j I_\Omega / (w_{Nij} T_{ij})$ if $w_{Nij} T_{ij} / p_j \leq I_\Omega$, and $w_{\Omega ij} = 1$ otherwise. Consequently,

$$\begin{aligned} w_{Sij} &= I_\Omega / T_{ij} \text{ if } w_{Nij} T_{ij} / p_j \leq I_\Omega \\ &= w_{Nij} / p_j \text{ if } w_{Nij} T_{ij} / p_j > I_\Omega \end{aligned} \tag{7}$$

Thus, for those units for which $w_{Nij} T_{ij} / p_j \leq I_\Omega$ we have that w_{Sij} is inversely proportional to T_{ij}.

If we had selected the entire S sample directly from C' without the intermediate step of selecting the NCS sample from $C' \cap G$, then the measure of size for each unit ij in C' for this sampling would be T_{ij} / p_j instead of $w_{Nij} T_{ij} / p_j$, the measure of size when selecting S from C''. The use of T_{ij} / p_j as a measure of size in this situation yield weights that are as close to being inversely proportional to T_{ij} as possible, using the reasoning above but with $w_{Nij} = 1$ for all units.

In fact, we first developed the measure of size T_{ij} / p_j as an appropriate measure of size for use when the sampling frame is a universe frame for the sample PSUs and then generalized this measure of size to apply to a situation when the final sample is selected from a frame like C'' which involves an intermediate stage of sampling.

4. Alternative Approaches to Sample Allocation and Selection

Before deciding on T_{ij} / p_j as the measure of size to use in the case of a universe frame, two alternatives to this measure of size were considered and rejected. Both of these alternative approaches would have allocated the total sample first among the sample PSUs and then selected the units within each PSU. The first would have allocated the total sample among the PSUs proportional to total employment in the PSU and then allocated within a PSU to each industry stratum proportional to the employment in the industry within the PSU. There are two problems with this approach. First, since a small PSU has a smaller probability of selection then a larger PSU, allocating a smaller number of establishments to such a PSU would result in a lower overall probability of selection, and hence a higher weight, for an establishment in a small PSU than for an establishment with the same assigned employment in a larger PSU. Secondly, if the set of sample PSUs were underrepresented in a particular industry, then the sample allocation to that industry would be low, increasing the variance of estimates for the industry.

The second alternative approach would have allocated among the sample PSUs proportional to the total employment in the geographic stratum, rather than the employment in the PSU, and then would have allocated among the industries within the PSU proportional to the employment in the industry in the geographic stratum, rather than the PSU. This approach has neither of the problems that the first approach does. However, it has the problem that if a PSU has a low proportion of its employment in a particular industry relative to the other PSUs in the geographic stratum, then the probability of selection for an establishment in that industry in the PSU would tend to be higher than for an establishment of the same size in a sample PSU in a different stratum that has a high proportion of its

employment in that industry relative to the other PSUs in the geographic stratum. In fact, if there are few enough establishments in an industry in the PSU, it is possible that the sample allocation to the PSU for that industry might be more than the number of establishments in the industry universe for that PSU.

The procedure of assigning a measure of size T_{ij}/p_j to each establishment and then selecting in each industry a systematic pps sample across PSUs is roughly equivalent to a third alternative, that is allocating separately in each industry among all PSUs proportional to the total employment within the industry within the PSU divided by the probability of selection of the PSU. There is one difference between the third alternative and the approach of assigning a measure of size T_{ij}/p_j to each establishment. If the sample in each industry stratum is first allocated among the PSUs and then independently selected in each PSU, as in the third alternative, then a noncertainty establishment with a given assigned employment would have a higher selection probability in a PSU for which a large proportion of the employment in the industry is in certainty establishments than would an establishment of the same size in a PSU with a small proportion of total employment in the industry in certainty establishments. This is because when the certainty establishments are removed and the sampling interval recomputed in the sampling process, this is done separately in each PSU in the third alternative.

5. Selection and Weighting of Subsamples of the ECI Samples

We now turn to the selection of the five subsamples $S_1,...,S_5$ discussed earlier and the selection and the weighting of the five final ECI samples $S_1',...,S_5'$. For any establishment ij in C_k, we let w_{ijk} denote the weight associated with selection of the units in S_k' taking into account all stages of selection. Then $w_{ijk}=0$ for ij in $C_k \sim S_k'$. Now if ij is in $C^* \cap G$ and, conditioned on Ω, it a certainty unit in selecting S from C^*, then, as noted earlier, the establishment is in each of the S_k, S_k' and hence by (7)

$$w_{ijk} = w_{Sij} = w_{Nji}/p_j. \tag{8}$$

The remaining units in S, which we denote by S^*, are sorted by PSU first and then within PSU by ECI PSIC and assigned employment. These units are allocated among $S_1,...,S_5$ as follows. Let n be the total number of such units and let n_k be the number of such units allocated to S_k, $k=1,...,5$. Typically, each n_k will be within 1 of $n/5$, although the procedure allows for any allocation. A systematic equal probability sample of S^* of size n_1 is selected and designated as S_1. A systematic equal probability sample of $S^* \sim S_1$ of size n_2 is next selected and designated as S_2 and so forth.

All units ij in $S_k \cap S^* \cap G$ are in S_k', $k=1,...,5$ and all units in $S_1 \cap S^* \cap L$ are in S_1', in accordance with our earlier description of the selection process. Consequently, for those units, by (7),

$$w_{ijk} = w_{Sij} n/n_k = I_\Omega n/(T_{ij} n_k), \tag{9}$$

where in (9), w_{Sij} is the weight associated with the selection of the S sample from C and n/n_k is the weight associated with selecting $S_k \cap S^*$ from S^*.

For $k=2,3,4,5$, the sampling and weighting is more complicated for selecting the units in $S_k' \sim G$, since, as mentioned earlier, C^* is used in obtaining the allocation of these units among the sample PSUs and PSICs, while the specific units selected are obtained from L_k', not L. That is, for each sample PSU j and PSIC c we let n_{jck} be the number of units in $S_k \cap L$ in that PSU × PSIC cell and then select a systematic pps sample of size n_{jck} from among units in L_k' that are in PSU j and PSIC c, where the measure of size for unit ij in L_k' is its assigned employment in L_k', denoted T_{ijk}. The selected units are the establishments in $S_k' \cap L_k'$.

To obtain w_{ijk} for units ij in L_k', we first let w_{ijk}' denote for a unit in S_k' the reciprocal of the probability of selecting ij from L_k' to be in S_k' conditioned on n_{jck}, and let $w_{ijk}'=0$ if $ij \in L_k' \sim S_k'$, where through the remainder of this section it is understood that c is the PSIC to which establishment ij belongs.

Then

$$E(w_{ijk}' | n_{jck}) = 1 \text{ if } n_{jck} \neq 0. \tag{10}$$

Consequently, if we let for ij in L_k',

$$w_{ijk} = \frac{1}{p_j} \frac{n_{jck}}{E(n_{jck}|\Omega)} w_{ijk}' \text{ if } j \in F, \quad = 0 \text{ if } j \notin F \tag{11}$$

it follows from (10) and (11) that $E(w_{ijk}|\Omega) = 1/p_j$ if $j \in F$, $E(w_{ijk}|\Omega) = 0$ if $j \notin F$, and hence $E(w_{ijk}) = 1$ as required to produce unbiased estimates of totals.

Next we compute $E(n_{jck}|\Omega)$ for $j \in F$ to completely evaluate (11). Let T_{jc} denote the total employment in PSU × PSIC cell jc on the L frame and n_{jc+} denote the number of units in $S \cap L$ from this cell. Let n'_{jc} denote the number of these units which, conditioned on Ω, are certainty unit in selecting S from C^* and let T'_{jc} denote the total employment in the cell among units on the L frame other than these n'_{jc} units. Then the total measure of size of all L frame establishments in cell jc other than the n'_{jc} conditional certainty units for the selection of S from C^* is T'_{jc}/p_j and consequently,

$$E(n_{jck}|\Omega) = \frac{n_k}{n_+} E(n_{jc+}|\Omega) = \frac{n_k}{n_+}\left(n'_{jc} + \frac{T'_{jc}}{p_j I_\Omega}\right) \quad (12)$$

Finally we combine (11) and (12) to conclude

$$w_{ijk} = \frac{n_{jck} n_+ I_\Omega w'_{ijk}}{p_j n_k \left(n'_{jc} + \frac{T'_{jc}}{p_j I_\Omega}\right)} \text{ if } ij \in S'_k \cap L'_k \quad (13)$$

Thus for ij in S'_k, w_{ijk} is given by (8), (9), or (13) depending on whether ij is in $S \sim S^*$, $S^* \cap G$, or L'_k.

Note that it may seem surprising that any units in $S \cap L$ can be conditional certainty units in selecting S from C^*, since the assigned employment of each such unit is less than 50. However, the measure of size for a unit used in selecting S from C^* is its assigned employment divided by the probability of selection of its PSU. In some cases the probability of selection of the PSU is so small that the measure of size is large despite the relatively small assigned employment.

REFERENCES

Black, S. R., Ernst, L. R., and Tehonica, J. (1997), "Sample Design and Estimation for the National Compensation Survey," in *Proceedings of the Survey Research Methods Section,* American Statistical Association, to appear.

BLS Handbook of Methods (Bulletin 2490, April 1997), Washington, D.C.: Bureau of Labor Statistics, pp. 57-69.

Ernst, L. R. (1989), "Weighting Issues for Longitudinal Household and Family Estimates," in *Panel Surveys*, eds. D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, New York: John Wiley & Sons, pp. 139-159.

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