SAMPLE DESIGN AND ESTIMATION FOR THE NATIONAL COMPENSATION SURVEY

Susan R. Black, Lawrence R. Ernst, Jason Tehonica, Bureau of Labor Statistics
Jason Tehonica, 2 Massachusetts Ave. N.E., Room 3160, Washington, DC 20212

Key Words: OCSP, ECI, EBS, Nonresponse adjustment, Variance estimation

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 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 ECI survey by producing national indexes which track quarterly changes in labor costs, and also cost level information annually on the cost per hour worked of each component of compensation.

The most important difference between the new NCS program and the old OCS program is that the OCS program used a fixed list of occupations for which compensation data were collected from all sampled establishments, thus publishing data for only a limited number of occupations. The NCS uses a probability selection of occupations in each establishment to insure a nearly universal coverage of occupations in the workforce. The NCS will be able to publish estimates for a greater number of occupations, as well as produce estimates for occupational groups. A second difference is that the NCS includes establishments with at least one worker, while the OCS program only used establishments with 50 or more workers.

While there are changes to the ECI arising from its integration with the NCS program, the overall effect on the ECI will be minimal in terms of the estimates produced. The main design difference is that the sample establishments for the ECI, as well as the EBS, will now be a subsample of the parent NCS establishment sample, which is drawn from a probability sample of metropolitan statistical areas and non-metropolitan counties. Previously the ECI and EBS sample establishments were selected from all in-scope establishments in the United States, without geographic clustering. This paper focuses mainly on the larger parent NCS sample.

This paper will describe the sample design and estimation process for the NCS. Section 2 covers the three stages of sample selection: the area based PSUs, the establishments, and the occupations. Section 3 will explain the weights associated with each stage of the sample design. This weighting discussion will include a new weight at the occupation level which will produce estimates reflecting current employment, instead of a weight that reflects employment at the time of initiation, like the weight currently used in the ECI.

Due to space limitations, two sections of the paper have been omitted, one which describes a method of allocating the establishment sample in each industry stratum among noncertainty PSUs which may be used in the future, and the other which presents a brief description of the variance estimation method used for the NCS. The complete paper is available from the authors.

2. Sample Selection

PSU Selection

The design of the NCS involves three stages of sample selection. The primary sample units (PSUs) are metropolitan statistical areas (MSAs) and non-metropolitan counties. The NCS sample PSUs are those originally chosen for the most recent design of the OCS program. The PSUs are broken down into three categories: certainty metropolitan areas, noncertainty metropolitan areas, and non-metropolitan counties. The certainty metropolitan areas are the Consolidated Metropolitan Statistical Areas (CMSAs), areas like New York, NY and Los Angeles, CA, and other MSAs with a total, non-agricultural employment greater than 560,000. This cutoff was chosen because those areas with employment totals greater than 560,000 generally have a significant federal work force and are of primary interest to the President's Pay Agent. In addition 3 areas, Raleigh-Durham, NC, Dayton-Springfield, OH, and Huntsville, AL, are certainty metropolitan areas because of the large federal work force in each of these areas, even though the total employment in each of these areas is below the 560,000 cutoff. The remaining metropolitan areas and the non-metropolitan counties were divided into MSA and non-MSA strata and then partitioned into regions using the Regional Classification of States from the Bureau of the Census. Within each region and type of area, the areas were ordered by average annual pay, and then strata were formed with approximately equal employment. The total sample of areas was allocated across the strata and then one area was selected in each stratum with pps using 12/92 employment numbers. Overall, there are 31 certainty met areas, 45 noncertainty met areas, and 70 non-met counties.

Establishment Selection

The second stage of sample selection is a set of establishments. An establishment is typically defined as a single physical location, although it sometimes consists of multiple locations of a company. An establishment is based on a single Universal Database (UDB) number. The UDB is compiled from lists of establishments from each state's unemployment insurance records. Each UDB number corresponds to a reporting number from the unemployment insurance database. The establishment, or single UDB number, is the assigned sample unit. However, government establishments can be defined differently due to the government clustering process used in creating the universe for the NCS. Government establishments are defined as a cluster of single physical locations with the same Employer Identification Number (EIN), which is an identifier on the unemployment insurance records used for IRS reporting purposes. This is necessary due to the fact that governments around the country are not consistent in their reporting techniques. The clustering process allows a uniform establishment definition for government units across geographical survey areas.

The sample of establishments within a geographical survey area is allocated by industry. The industry strata were chosen in a way as to produce estimates for each Major Industry Division, like Retail Trade, as well as selected, more narrowly defined industries which are traditionally produced for the ECI, like General Merchandise Stores and Food Stores.

In certainty areas the total number of establishments to be selected is determined by local area publication needs. The total number of sample establishments to be selected for noncertainty areas is allocated among the geographic strata proportional to the geographic stratum size. In each PSU, the allocation of establishments among the industry strata is proportional to size, with the constraint of a pre-defined minimum of establishments in each industry stratum. Allocating the sample proportional to size among the industry strata results in a lower variance at the aggregate level while instituting a minimum sample for each stratum lowers the variance at the individual industry breaks. The minimum sample for the industry strata varies between certainty areas and noncertainty areas. The minimum is lower for noncertainty areas because the focus is not on producing estimates for each industry at the PSU level. Instead the noncertainty allocations are designed to produce estimates at national and regional levels. Within each industry stratum the designated number of sample establishments are selected pps.

Occupational Selection

The third and final stage of selection is a set of occupations. Within each establishment, a sample of "defined occupations" is chosen by probability selection of occupations (PSO). For NCS purposes, a defined occupation is based on the occupation classification of the Census of Population, further defined by the following characterizations: full-time or part-time, union or nonunion, time or incentive, and a determination of the level of work. The occupation's Level is determined using a generic leveling process which ranks and compares all occupations selected in an establishment using the same criteria. This is a departure from the OCS program which used leveling definitions unique to each occupation. Refer to Cohen (1997) for more information on the generic leveling process. The number of defined occupations (quotes) sampled within each establishment is a function of the PSO employment of the establishment. This employment is determined by creating a list of all occupations in the establishment eligible for PSO, which at present does not include certain types of occupations, such as those where the worker sets his or her own pay. Refer to Cohen (1997) for more information on the number of quotes selected in each establishment.

Once a list of eligible occupations is assembled and the number of quotes needed is determined, the quotes are selected by an equal probability systematic sample of individual employees in the establishment. This is equivalent to a pps sample of the actual defined occupations, because once an employee is selected, data is collected for all employees in that establishment with the same defined occupation.

In some instances, the collected data may not match exactly the assigned establishment, for example, a company may be unwilling or unable to separate out data for a single location and instead can only provide data for more than one UDB number. Weighting adjustments are made to account for any collected data that does not match the assigned unit. Other complications can arise at the occupation level, such as subsampling down to a particular department within an establishment. For more details, refer to Black, Ernst, and Tehonica (1997).

3. Weighting

An independent sample is drawn for each PSU, resulting in separate weighting for each area sample. The description of the weights in the following section reflect only the sampling of establishments and occupations for locality estimates. The weights are multiplied by the reciprocal of the probability of selecting the corresponding PSU for national, regional, and "Rest of US" estimates. To obtain a final establishment weight, the assigned sampled establishment weight, which is the reciprocal of the probability of selecting the establishment, is multiplied by various establishment level adjustment factors. The adjusted establishment weight is then multiplied by the PSO sampling interval and occupation level adjustment factors to obtain a final occupation weight. The occupation weight does not reflect the reciprocal of the probability of selecting the specific occupation, but must be divided by the total number of employees in the selected occupation to reflect this value. This results in another final weight known as the employee or individual weight. It is the employee weight, rather than the occupation weight, that is used in all NCS estimates. Note that the NCS does not produce any establishment based estimates, and therefore the final establishment weight is not directly used in any NCS estimates.

Establishment Weighting

The establishment weighting process begins with the assigned establishment weight which is simply the reciprocal of the probability of selecting the establishment from the UDB given that the establishment's PSU has been selected.

The next stage of the weighting process at the establishment level, is the establishment nonresponse adjustment which takes a weighting cell approach. All establishments are put into a defined nonresponse adjustment cell based on assigned employment, assigned Standard Industrial Classification (SIC) code, and assigned sector, that is, private industry, local government, or state government. When an establishment is considered a refusal nonrespondent, its assigned weight multiplied by its employment is redistributed to the respondents in the same nonresponse adjustment cell. This is done by calculating a nonresponse adjustment factor that is applied to the weight of all responding establishments.

Let AW_{ki} be the assigned establishment weight for establishment i in nonresponse cell k and let E'_{ki} be the assigned employment for this establishment. The number of sample units, both respondents and nonrespondents, in nonresponse cell k is denoted as T_k of which the first R_k are the responding units. The nonresponse adjustment factor, F_k, that is applied to the responding units in cell k is:

$$F_k = \frac{\sum_{i=1}^{T_k} AW_{ki}E'_{ki}}{\sum_{i=1}^{R_k} AW_{ki}E'_{ki}} .$$

Note that this factor redistributes the assigned sample employment and thus preserves the weighted employment total in each cell which, by the method of sampling used in the NCS, is always nearly equal to the frame employment total for that cell. If instead of computing F_k as we have done, we had omitted E'_{ki} in the above formula, we would have preserved the weighted total number of establishments in each cell rather than the weighted employment total. Since all NCS estimates are employee based it is more important to preserve employment totals.

There are additional establishment level adjustments that occur when the collected unit differs from the originally assigned unit. The different types of situations where such adjustments are needed are listed below:

Subsampling of physical locations: (SAF_{ki})

Merges: (MAF_{ki})

Collected more or less than assigned: (DAF_{ki}).

These adjustments and how they are computed, are described in detail in Black, Ernst, and Tehonica (1997).

The final establishment weight, FW_{ki}, is computed by multiplying the assigned sampled establishment weight AW_{ki}, which is the reciprocal of the probability of selecting the establishment, by the adjustment factors for nonresponse, documentation, and merges.

$$FW_{ki} = AW_{ki} \times F_k \times DAF_{ki} \times MAF_{ki} \times SAF_{ki}$$

An additional weighting step is needed in a merge situation. The final weights computed as above for each assigned unit that is part of the merge are summed to obtain a combined weight that applies to the combined data of all locations that the company is reporting for. This step is described in detail in Black, Ernst, and Tehonica (1997).

Occupational Weighting

For occupational weighting for occupation j in establishment i, we begin with the final establishment weight for establishment i, which we now denote as FW_i, dropping the subscript for the establishment nonresponse adjustment cell. We multiply this weight by the PSO sampling interval, I_i, used in the occupational selections for establishment i. The PSO sampling interval is the number of PSO employees in the establishment divided by the number of occupational selections. This is then multiplied by the duplication or collapsing factor, C_{ij}, that designates how many times occupation j in establishment i was selected during PSO. The product of these three terms is referred to as the occupation weight before occupational nonresponse adjustment, which we denote by OW_{ij}.

$$OW_{ij} = FW_i \cdot I_i \cdot C_{ij}$$

Note that I_i is not the reciprocal of the probability of selecting occupation j in establishment i given that that the establishment has been selected. To obtain that quantity we would have to divide I_i by the number of employees in establishment i that have occupation j. This will be discussed later in more detail.

Two tiers of occupational nonresponse adjustments are applied to the occupation weight, OW_{ij}, to obtain the final occupation weight. The first is associated with the Level of the occupation, the second with the Major Occupational Group (MOG) of the occupation. Each occupation sampled is coded into a MOG and is leveled using generic leveling criteria as described earlier. The nonresponse adjustment cells for the Level tier of adjustment are defined by the establishment's reported industry, reported employment size, and the occupation's MOG and Level. For the MOG tier of adjustment, the adjustment cells are defined by the establishment's reported industry, reported employment size, and the occupation's MOG.

The need for these two tiers of adjustment is explained below. During collection there are times when the field economist will not be able to obtain the desired collection data for a particular sampled occupation from a respondent establishment. When this happens, the occupation is deemed a nonrespondent and an adjustment is made. Two of the four variables that are used in forming the nonresponse adjustment cells, reported industry and reported employment, are always known for a respondent establishment with a nonrespondent for a particular occupation. A requirement for data collection is that a third variable, MOG, be obtained for all occupations, including nonrespondent occupations. The fourth variable, the Level of the occupation, may or may not be obtained for a nonrespondent occupation. If the Level information is obtained, the nonrespondent occupation is assigned to the adjustment cell for the Level tier of adjustment to which it belongs. If the Level information cannot be obtained, the occupation is not used during the Level tier of weight adjustment. It is assigned to the appropriate adjustment cell for the MOG tier of adjustment.

It is worth noting that there are occupations that are not able to be leveled. These are occupations such as artists, dancers, and actors to which a Level is not appropriate. Occupations of this type are put into a separate "Level" of their own and are used during the Level tier of adjustment.

The level tier of adjustment is performed first. The adjustment factor for the k-th nonresponse cell for the level tier is denoted as F_{Lk} and computed as follows. Let R_{Lk} be the set of respondent occupations in this cell and let T_{Lk} be the set of all occupations, respondents plus Level tier nonrespondents in the cell. Then:

$$F_{Lk} = \frac{\sum_{ij \in T_{Lk}} OW_{ij}}{\sum_{ij \in R_{Lk}} OW_{ij}}$$

The occupational weight for occupation ij, after the Level tier nonresponse adjustment, is denoted as LOW_{ij}. For a respondent occupation it is

$OW_{ij} \cdot F_{Lk_{ij}}$ where k_{ij} is the level tier nonresponse adjustment cell for occupation *ij*. For a MOG tier nonrespondent occupation, $LOW_{ij} = OW_{ij}$, that is no adjustment is applied to these nonrespondents.

The MOG tier of adjustment is then performed. It is computed similarly to the Level tier adjustment with LOW_{ij} replacing OW_{ij}. The adjustment factor for the *k*-th nonresponse cell for the MOG tier is denoted as F_{Mk}. Now R_{Mk} is the set of respondent occupations in this cell and T_{Mk} is the set of all occupations, respondents plus MOG tier nonrespondents, in the cell. Then

$$F_{Mk} = \frac{\sum\limits_{ij \in T_{Mk}} LOW_{ij}}{\sum\limits_{ij \in R_{Mk}} LOW_{ij}}.$$

The final occupational weight for respondent occupation *ij*, which is obtained after the MOG tier nonresponse adjustment factor is applied, is denoted as FOW_{ij}. It is simply $LOW_{ij} \cdot F_{Mk_{ij}}$ where k_{ij} is the MOG tier nonresponse adjustment cell for occupation *ij*.

Thus the final occupation weight for occupation *ij*, is the final establishment weight for establishment *i* multiplied by the PSO sampling interval, the duplication factor, and the two occupation nonresponse adjustment factors.

One more weight is needed. The final employee or individual weight, is denoted by EW_{ij}, for occupation *j* in establishment *i*. EW_{ij} is simply defined as

$$EW_{ij} = FOW_{ij} / N_{Fij}$$

where N_{Fij} is the number of PSO employees in occupation *ij* at the time of the first or initiation interview. It is EW_{ij} rather than FOW_{ij} that is used to obtain all NCS estimates.

Much of the weighting for the NCS reflects the approach used in the ECI, since the ECI concept of first sampling establishments and then sampling defined occupations from an establishment list of employees or occupations for each sample establishment has been carried over to the NCS. One major difference is that the ECI has been using a weight essentially equivalent to the occupation weight defined above, which also is known as the "quote weight," to compute estimates of mean wages. NCS is using the employee weight defined above, for estimates of mean wages, total number of workers in a domain and for positional statistics, such as medians. Generally these two weighting approaches yield precisely the same estimates of mean wages and total workers at the time of the initiation interview. However, for update interviews, estimates of means produced using quote weights, unlike those produced using employee weights, only reflect the employment at the time of the initiation interview, not current employment. Furthermore, quote weights are completely inappropriate for computing positional statistics. We first proceed to describe these two approaches to weighting. The description focuses only on the components of these weights arising from the probability selection of the establishment and occupations. All the components relating to nonresponse and other adjustments are ignored. Therefore, we will obtain a somewhat simplified version of the occupation weight and employee weight that we have previously defined and will use a different notation.

For the *i*-th establishment let: p_i be the probability of selection of the establishment; I_i be the sampling interval used in the occupational selections, that is the number of employees on the sampling list divided by the number of occupational selections; and N_{Fij}, N_{Cij} denote the number of employees in occupation *j* in this establishment at the time of the first, or initiation interview, and at the time of the current interview, respectively. Then the quote weight w_{Qij} for occupation *ij* is I_i / p_i if both the establishment and occupation are selected; otherwise $w_{Qij} = 0$. Similarly the employee weight w_{Eij} for occupation *ij* is $I_i / (N_{Fij} p_i)$ if both the establishment and occupation are selected; otherwise $w_{Eij} = 0$.

An estimate, for example, of total wages in a domain using employee weights is obtained by multiplying the employee weight for each current employee within the domain by the employee's wages and summing the product over all sampled employees within the domain. An estimate of total employment within a domain is obtained by summing the employee weights over all current sample employees within the domain. An estimate of mean wages using employee weights is obtained by taking the quotient of the previous two estimates.

An estimate of total wages in a domain using quote weights is obtained by multiplying the quote weight for each quote within the domain by the mean wages for the quote and summing the product over all quotes within the domain. An estimate of total employment in a domain using quote weights is obtained by summing the quote weights over all quotes within the domain. An estimate of mean wages using quote weights is obtained by taking the quotient of the previous two estimates.

To establish that employee weights yield unbiased linear estimators of the data, such as total employees or total wages in a domain, while the quotes weights do not, we first observe that since p_i is the probability of selection of the establishment and N_{Fij} / l_i is the expected number of times that occupation ij is selected given establishment i is selected, it follows that

$$E(w_{Eij}) = 1 \qquad (1)$$

From (1) and the definitions of the employee and quote weights it follows that

$$E(w_{Qij}) = N_{Fij} . \qquad (2)$$

If Y_{Cijk} is the value of the characteristic of interest for the employee k in occupation ij at the current time, $\overline{Y}_{Cij} = \sum_{k} Y_{Cijk} / N_{Cij}$. If Y_C is the population total and $\hat{Y}_{EC}, \hat{Y}_{QC}$ are the employee weight and quote weight estimators of Y_C, respectively, then by (1) and (2),

$$E(\hat{Y}_{EC}) = \sum_{ijk} E(w_{Eij}) Y_{Cijk} = \sum_{ijk} Y_{Cijk} = Y_C = \sum_{ij} N_{Cij} \overline{Y}_{Cij} \qquad (3)$$

$$E(\hat{Y}_{QC}) = \sum_{ij} E(w_{Qij}) \overline{Y}_{Cij} = \sum_{ij} N_{Fij} \overline{Y}_{Cij} \neq Y_C . \qquad (4)$$

While by (3), $\hat{Y}_{EC}$ is an unbiased estimator of Y_C, $\hat{Y}_{QC}$ is a biased estimator by (4) that weights the mean $\overline{Y}_{ijC}$ by the employment N_{Fij} at the time of the initiation interview rather than the current N_{Cij}.

Note that in the special case when the characteristic of interest is total employment in a domain D, then the characteristic value for each employee associated with a quote is 1 and therefore $\overline{Y}_{Cij} = 1$ for all ij. Then, by (3), (4), $E(\hat{Y}_{EC}) = \sum_{i,j \in D} N_{Cij}$, $E(\hat{Y}_{QC}) = \sum_{i,j \in D} N_{Fij}$ and thus $\hat{Y}_{EC}, \hat{Y}_{QC}$ estimate total employment in D at the time of the current interview and the initiation interview, respectively.

The observation in the previous paragraph, that the sum of the quote weights in a domain estimates the employment in the domain at the time of initiation, explains the one use of quote weights in the occupational weighting process, namely in the occupational nonresponse adjustment. The occupational nonresponse adjustment redistributes the occupation weights of the nonrespondents in a cell at initiation to the respondents and therefore preserves the estimated employment in the cell. Employee weights could not be used for this purpose, because employee weights depend on the number of employees in an occupation, which is not known for nonrespondent occupations. This explains why the employee weight is not defined until the last step of the entire weighting process, subsequent to the occupational nonresponse adjustments.

References

Black, S. R., Ernst, L. R., and Tehonica, J. (1997), "Statistical Issues Associated with Data Collection for the National Compensation Survey," in *Proceedings of the Survey Research Methods Section,* American Statistical Association, to appear.

Cohen, S. H. (1997), "The National Compensation Survey: The New BLS Integrated Compensation Program," in *Proceedings of the Survey Research Methods Section,* American Statistical Association, to appear.

Paben, S. P., and Ernst, L. R. (1997), "Adjusting the Number of Occupational Selections for the National Compensation Survey." in *Proceedings of the Survey Research Methods Section,* American Statistical Association, to appear.

Tehonica, J., Ernst, L. R., and Ponikowski, C. H. (1997), "Summary of Estimation and Variance Specifications for the 1996 Albuquerque, NM COMP2000 Test Survey," Bureau of Labor Statistics memorandum to The Record, dated March 3.

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