ADJUSTING ESTABLISHMENT SELECTION PROBABILITIES AND NUMBER OF OCCUPATIONAL SELECTIONS TO REDUCE VARIANCES IN BLS COMPENSATION SURVEYS

Steven P. Paben, Lawrence R. Ernst, Bureau of Labor Statistics
Steven P. Paben, BLS, 2 Massachusetts Ave., N.E., Room 3160, Washington, DC 20212

Key Words: NCS, Sampling, Variance reduction

INTRODUCTION

The National Compensation Survey (NCS) is both a replacement for the Occupational Compensation Survey Program (OCSP), which measures employee wages by work level, and also a program that integrates this new survey with two existing BLS compensation surveys, the Employment Cost Index (ECI), and the Employee Benefits Survey (EBS). NCS uses a rotating panel design with three stages of selection used in selecting each panel: geographic area PSUs; establishments selected from industry strata; and occupations selected separately from each sample establishment. The establishments are selected with probability proportional to size (pps), with total employment the measure of size. For each selected establishment the number of sampled occupations are a nondecreasing function of employment size, with the occupations selected pps.

This study of the sample design has two purposes. First, to determine whether differential sampling intervals for establishments in different size classes would be more effective for the purpose of variance minimization than sampling of establishments by direct pps. Second, to determine whether a different allocation than the current allocation of the number of occupation selections within an establishment by size class can reduce variances without increasing the total number of occupational selections across all sample establishments.

SAMPLE DESIGN

Much of the focus in NCS, including all the analysis in this paper, is on the production of locality estimates, that is estimates for individual PSUs, particularly certainty PSUs, for which the first stage of sampling is not an issue. Consequently, this stage of sampling is not addressed here. In the second stage of sampling, establishments are selected pps from industry strata, with total employment the measure of size. The sampling frame from which the establishments are selected is constructed primarily from the unemployment insurance universe.

In the third stage of sampling, occupations are selected separately from each establishment. Typically, the occupational selections are done from a complete list of in scope employees for the establishment obtained from the respondent. (Certain cases of employees, such as those who set their own pay are out of scope). A systematic equal probability sample of employees is selected. Then, for each selected employee, wage data is obtained for all employees with the same detailed job as the selected employee within the particular establishment. For example, if one of the employees selected is a full time, grade 9, non-union, accountant, whose earnings are time based (as opposed to incentive based), then data is collected for all employees satisfying these criteria for that establishment. Consequently, the equal probability selection of employees is equivalent to a pps selection of detailed jobs. The number of occupational selections in each establishment is dependent upon the size of the establishment. For each of the surveys that we studied the number of occupational selections was determined by the following schedule:

Table 1.

Number of Employees	*Number of Selected Occupations*
0 - 49	4
50 - 99	8
100 - 249	10
250 - 499	12
500 - 999	16
1000 +	20

The weight for each employee in a selected job is obtained by taking the product of the reciprocal of the probability of selecting the establishment, the reciprocal of the probability of selecting the job given that the establishment is selected, and nonresponse adjustment factors for establishment and occupational nonresponse.

METHODS

In order to determine an optimal allocation, the NCS test variance program and procedures were used. We used data from five of the six Metropolitan Statistical Areas (MSAs) covered by this 1996 test survey program. One of the test surveys was not used due to some outliers in the original variance calculation. The optimization was based on the variance for weekly mean wage.

The variance to be minimized has a between and a within establishment component. The number of establishments sampled affects both components. The number of occupational selections affects only the within establishment component of the variance. Since our sample has multiple stages of selection, this problem could have been treated as a nonlinear optimization problem. However, we concluded it was somewhat simpler to view the problem as two Neyman allocations and that this approach would produce similar results, as we establish at the end of this section.

The first allocation to optimize is the number of establishments. We actually are trying to find an optimal allocation for differential sampling intervals based on the size classes shown in Table 1. The sampling interval for the *i*-th size class is obtained by dividing the overall sampling interval by a factor, f_i, and thus the problem is reduced to finding an optimal set of f_i's. This factor increases or decreases the allocation depending on whether it is greater or less than one. The constraint on this optimization is that the total number of establishments summed over the five test surveys and six size classes must remain the same.

The second allocation to optimize is the number of occupational selections within each sampled establishment. The current allocation uses the schedule defined in Table 1. We want to optimize this allocation using the same size classes. The constraint for this allocation is that the total number of occupational selections must remain the same.

The following method is similar for both allocations. We start with the optimal allocation for the differential sampling intervals. The objective function is

$$\sum_{i=1}^{6}\sum_{j=1}^{2}\frac{\hat{\gamma}_{Bij}^{2}}{f_i}, \qquad (1)$$

where $\hat{\gamma}_{Bij}^{2}$ is the between establishment component of the estimated relative variance for the *i*th size class and *j*th industry stratum summed over the five areas, using the original allocation of establishments. Only two industry strata were used in the test surveys, namely the government (state and local) and private sectors. Thus, it is not actually a single variance that we are minimizing but a sum of relative variances.

The relative variance is the variance divided by the mean wage squared. The between establishment component of the relative variance is the between establishment component of the variance divided by the mean squared, with an analogous definition for the within establishment component of the relative variance. The between establishment component of the relative variance was used since this component is not affected by the allocation of occupations. The effect of minimizing the between establishment component in this optimization will be seen at the end of this section.

The following optimization method is similar to the one due to Neyman described in Cochran (1977) for finding an optimum allocation in a stratified random sample. The linear objective function (1) subject to the constraint

$$\sum_{i=1}^{6} f_i n_i = n, \qquad (2)$$

on the total number of noncertainty establishments is minimized by

$$f_i = f\,\frac{\hat{\gamma}_{Bi}}{\sum_{i}\hat{\gamma}_{Bi}}, \qquad (3)$$

where,

$$f = \frac{n\sum \hat{\gamma}_{Bi}/\sqrt{n_i}}{\sum \hat{\gamma}_{Bi}\sqrt{n_i}}, \qquad (4)$$

n_i is the number of noncertainty establishments in each size class over all areas and strata, and n is the total number of noncertainty establishments.

We use a similar process in finding the optimal allocation for the number of occupational selections. Here, we use the optimal allocation of establishments obtained by the procedure just described and for this establishment allocation determine the occupational allocation that minimizes the within component of the relative variance. Both noncertainty and certainty establishments are used since they both contribute to this component of the relative variance. The linear objective function in this case is

$$\sum_{i=1}^{6}\sum_{j=1}^{2}\frac{\hat{\gamma}_{Wij}^{2}}{f_{ij}}\cdot\frac{m_i}{m_i^{*}}, \qquad (5)$$

where $\hat{\gamma}_{Wij}^{2}$ is the estimated within component of the relative variance for the *i*th size class and *j*th stratum summed over the five areas for the actual allocation of establishments and occupational selections used in the test surveys. (Actually this component of the relative variance contains a finite population correction term which is omitted from $\hat{\gamma}^2_{Wij}$ since it is not a function of the number of occupational selections and consequently does not affect the optimization.) $f_{ij} = f_i$ for noncertainty units and $f_{ij} = 1$ for certainty units, where the f_i's are now constants obtained from the first optimization. f_{ij} is present in (5) since the original number of noncertainty establishments in size class i has been multiplied by f_i. m_i is the actual number of occupational selections in each size class. m_i^* is the optimal allocation for the number of occupational selections in each size class, with the m_i^*'s the variables in the second optimization. The term m_i / m_i^* is present in (5) to adjust for the effect on the relative variance of selecting m_i^* occupations instead of m_i.

To obtain the optimal m_i^*'s we minimize (5) subject to the constraint

$$\sum_{i=1}^{6} m_i^*(f_i n_i + n_i') = C, \qquad (6)$$

where n_i, n_i' are, respectively, the number of noncertainty and certainty establishments in each size class for all areas and strata in the original allocation, and C is the total number of occupational selections. The optimal m_i^*'s are obtained from formulas analogous to (3) and (4) for the first optimization.

We now investigate the question of whether the f_i^*'s and m_i^*'s obtained by the two optimizations would be the same as the optimal values obtained by minimizing both sets of variables simultaneously for the objective function (1)+(5) subject to constraints (2) and (6). ((1)+(5) is the appropriate objective function since it is the overall relative variance.)

In the special case when all establishments in all size classes are noncertainty, the simultaneous optimization approach will yield the same allocation as obtained with our two optimizations approach. To establish this claim we first observe that (1) + (5) reduces, with the substitution $x_i = f_i m_i^*$ in (5), to

$$\sum_{i=1}^{6}\sum_{j=1}^{2}\left(\frac{\hat{\gamma}^2_{Bij}}{f_i} + \frac{\hat{\gamma}^2_{Wij} m_i}{x_i}\right) \qquad (7)$$

and (6) reduces to

$$\sum_{i=1}^{6} x_i n_i = C, \qquad (8)$$

with (2) remaining unchanged.

Then, note that since f_i is only present in the first term in (7) and in (2), minimizing (7) subject to (2) and (8) yields the same f_i as minimizing (1) subject to (2). Furthermore, since the x_i's are present only in the second term in (7) and in (8), minimizing (7) subject to (2) and (8) yields x_i which when divided by the optimal f_i produce the same m_i^* as would be obtained by minimizing (5) subject to (6).

If there are certainty establishments, as there are in our applications then the two approaches do not generally yield the same allocations, and consequently, our approach using two sequential optimizations does not yield the optimal simultaneous allocation for both sets of variables.

Note if we were interested in finding only the optimum allocation for the number of establishments by size class instead of the optimal allocation of the number of establishments and occupations we would have minimized (1) subject to (2), with the between establishment component of the relative variance replaced by the overall relative variance. Likewise, to find the optimal allocation for the number of occupational selections alone, we would have minimized (5) subject to (6) with f_{ij} replaced by 1 in (5) and (6).

We have omitted here providing details of the calculations of the between and within establishment variance components used in (1) and (5). A complete description is provided in Tehonica, Ernst, and Ponikowski (1996). The following is a very brief overview. The variance estimation formulas for mean wages are obtained using a linearized Taylor Series form. For noncertainty establishments the overall variance is estimated with a pps with replacement formula reflecting the fact that the first stage of sampling in a PSU is a pps sample of establishments. The within establishment component of variance for both certainty and noncertainty establishments is estimated by a simple random sample without replacement formula, reflecting the fact that the sample of occupations within an establishment is typically obtained through a systematic sample of employees. The between establishment variance estimate is obtained by subtracting the within establishment variance estimate for noncertainty establishments from the total variance estimate for noncertainty establishments.

An interesting feature of the objective functions (1) and (5) is that they allowed us to find and see the effect of the optimal allocation, while using the original variance formulas. The one exception is that we did not use the fpc term of the within establishment variance in determining the optimal allocation for the second optimization. However, the fpc term was used when comparing the variance with the optimal allocation against the variance with the original allocation.

RESULTS

We are primarily interested in the effect the sequential optimal allocations have on the variances for mean wages in three domains: overall, Major Occupational Groups (MOGs), and work levels. Each of these domains and each group or level within each domain calls for a different optimal allocation. Since only one allocation is possible we use an allocation that is a weighted average of the allocations that are optimal for these characteristics.

The overall domain simply includes all occupational selections. The MOGs domain groups occupational selections into 10 categories, such as Professional, Sales, and Service Occupations. The work levels domain groups occupational selections into 15 different levels depending on the total score of the selection on ten different factors, such as knowledge required, guidelines given, and supervisory duties. The work levels are analogous to the federal government's GS levels.

The adjusted allocation that we used was obtained as follows. An optimal allocation was obtained for the sum of the relative variances for overall mean wages over the five test surveys. Ten additional optimal allocations were obtained, each for the sum of the relative variances of the mean wages for a specific MOG. The final allocation was a weighted average of these 11 allocations, with each MOG allocation equally weighted and the overall allocation given a weight 10 times larger than the weight of each MOG allocation. It was decided not to base the allocations on the work levels, since there were few observations at the higher work levels in some of the size classes. We did not want the allocations to be overly influenced by a handful of observations. However, we did observe the effect the adjusted allocation had on the levels.

Note that an alternative allocation could have been obtained by using the using an objective function that is a linear combination of the 11 objective functions actually used with the relative variance overall being given 10 times larger weight in this objective function then the relative variance for each MOG. That is, this alternative uses a weighted average of the objective functions instead of a weighted average of the allocations. With this approach the characteristics with larger relative variances would influence the allocation more than the other characteristics. This may be either a drawback or an advantage, depending on whether there is more interest on improving the relative variance of all characteristics or just those with larger relative variances. If the former is the case then the linear combination of the 11 objective functions just described can be modified by dividing the weight for each characteristic by the sum over the five areas of the relative variances of that characteristic with the current allocations. With this modification we would be optimizing a weighted percentage reduction in the relative variances, summed over the five areas, of the 11 characteristics.

For the adjusted allocation that we ended up using, the factors, f_i, which determines the sampling intervals and the number of occupational selections are as follows:

Table 2.

Size Class	f_i	*Original Occ. Selections*	*Adjusted Occ. Selections*
0 - 49	1.0818	4	6
50 - 99	0.8649	8	6
100 - 249	1.0127	10	8
250 - 499	1.0047	12	9
500 - 999	0.7478	16	8
1000 +	0.4460	20	31

At first glance, there appears to be an inconsistency between f_i and the average optimal number of selections in the largest size class. The factor says to decrease the number of establishments to .4460 of the original number of establishments which will reduce the number of occupational selections, but the adjusted number of occupational selections says to increase the number of selections to 31 per establishment. However, the factor affects noncertainty establishments only. This apparent inconsistency is a result of the fact that only 10 of 113 establishments with more than 1000 employees were noncertainty establishments. Since there were so few noncertainty establishments in this size class, little credence should be given to this factor. Note that because the factor only affects the few noncertainty establishments it has little influence on the optimal number of selections in this size class which is determined by both certainty and noncertainty establishments.

We did not find the small differences among the five smallest size classes in the adjusted allocation of the number of occupational selections and the dramatically larger allocation for the largest size class to be surprising. Typically, variances tend to be lowered when the weights are inversely proportional to size. In the case of an occupational selection the "size" is the number of employees in the occupation. The employee weight, ignoring adjustments, is the product of the establishment weight and the reciprocal of the probability of selecting the occupation given that the establishment has been selected. The first term in this product, the establishment weight, is for a noncertainty establishment in each industry inversely proportional to the establishment employment. The second term is the sampling interval used in the probability selection of occupations (known as the PSO sampling interval) divided by the number of employees in the occupation. The PSO sampling interval is the number of employees in the establishment divided by the number of occupational selections. (The establishment employment on the sampling frame used in selecting the establishment differs somewhat from the employment used in selecting the occupations, which is done at the time of the interview and which excludes certain classes of employees, but the difference is generally small and will be ignored in this discussion.) Consequently, the employee weight would be inversely proportional to the number of employees in an occupation if the product of the establishment weight and the PSO sampling interval is the same for all establishments. For noncertainty establishments this condition would be met if the number of occupational selections were the same in each of the size classes, since then the smaller establishment weight and the larger PSO sampling interval for larger establishments would cancel each other out. However, once the establishment employment goes above the certainty cutoff we would no longer have this cancellation. That is, as the establishment size continues to increase the establishment weight remains fixed at 1, but the PSO interval continues to increase unless the number of occupational selections is increased. Thus, we believe that the reason the allocation is so large for the largest size class is that most of the establishments in this size class are well above the certainty cutoff.

Since the factors for the differential sampling intervals are close to one, except in the two highest size classes where there are not many noncertainty establishments, there was little impact on the between establishment variance from this part of the allocation. However, the adjusted allocation for the number of occupational selections had a definite impact on the within variance, as can be ascertained from Table 3.

The estimates in Table 3 were obtained as follows. We first computed the estimate of each component of the variance (total, within, and between variance) for the overall domain and each MOG in the five test surveys using the adjusted allocation, and for each of the 55 estimates compared it to the original component of the variance by taking a ratio of the estimated adjusted variance component over the original variance component. We then computed the weighted average of the ratios for each test survey with the MOGs equally weighted and the overall domain weighted ten times any individual MOG. Finally, we found the grand weighted average of the ratios for all five test surveys combined as the arithmetic mean of the five weighted averages for the individual surveys, which is shown in Table 3.

Table 3.

Grand avg. of weighted total ratios	= 0.9412
Grand avg. of weighted within ratios	= 0.8428
Grand avg. of weighted between ratios	= 0.9969

That is, on the average, the estimated total variance decreased by about 6%, the estimated within variance decreased by about 16%, and the estimated between variance remained about the same using the adjusted allocation.

Variance estimates for 11 characteristics in each of five areas were used in determining the optimal allocations. The within ratio for 45 of these 55 characteristics was numerically less than 1, a striking result. If these 55 ratios were all independent estimates (which they are not) with each estimates having probability 1/2 of being less than 1, then the probability of obtaining so many ratios less than 1 would be negligible. That is, by the sign test the large number of ratios numerically less than 1 is significant at any reasonable level of significance. The within ratio for the overall estimate was less than 1 for each of these five areas. Even though one of the test surveys, the Salt Lake City MSA, was not used in determining the allocation, 8 of the 11 within ratios were less than 1 for this area.

Even though the variances of the work levels were not taken into consideration in finding our adjusted allocation, the adjusted allocation does quite well in reducing the variances. In a similar fashion as the calculations in Table 3, we found the grand average of the ratios of the adjusted variance over the original variance for each component of the variance over the

15 work levels for the 5 test surveys with each of the work levels equally weighted. The grand average of the total ratios was 0.9320. The grand average of the within ratios was 0.8880. The grand average of the between ratios was again very close to 1. Here, there were variance estimates for 75 characteristics from 15 work levels in five areas. The within ratio for 52 of these 75 characteristics was less than 1. Again, if we consider these ratios to be independent with each estimate having the probability 1/2 of being less than one, then the probability of obtaining so many ratios less than one would be small. As for Salt Lake City survey, 10 of the 15 within ratios were less than one.

The average ratios of the levels for the five test surveys are less than one for 13 of the 15 levels, which by the sign test is significant at the level of significance $\alpha = .01$. The average ratio is greater than 1 for work levels 14 and 15. Because of the small amount of data at these levels they happen to be the two levels for which there is the most interest in decreasing the variance. However, the increase in the variance estimates for these levels may be more a result of the limitations in the data than in the adjusted allocation. For example, in the Rochester test survey for work level 15, which had the greatest percentage increase in the within variance from the original to the adjusted allocation among all areas, only about 0.4% of all occupational selections are at this level, and none of them are in the highest size class where the adjusted allocation has the greatest percentage increase in occupational selections. This is a case where we believe the sample does not provide an estimate of the distribution by employment class that is close to the population distribution, since we do not believe that there are no level 15 employees in Rochester in establishments in the largest size class. There are similar explanations, although not as dramatic, for some of the other increases in the within variance for levels 14 and 15 with the adjusted allocation. Given the sparse amount of data at the levels, we believe we need data from additional areas to properly assess the effect of the adjusted allocation on variances for these levels.

CONCLUSIONS

In the NCS variance calculation for the locality mean wage estimates, there was only a small difference in the between component of the relative variance using differential sampling intervals compared to the between component of the original direct pps allocation. The results we have obtained do not provide much support for deviating from direct pps in selecting establishments. However, the within component of the relative variance was quite different from the optimal number of occupational selections. This allocation shows that for variance purposes it may be more appropriate to use only one or two different selection rates for establishments with less than 1000 employees, and that there should be an increase in the number of occupational selections for establishments with more than 1000 employees. This analysis does not include constraints imposed by respondent burden considerations. For example, regardless of the effect on variances, it is likely that an increase in the number of occupational selections for the smallest size class would not be given serious consideration for reasons of respondent burden.

REFERENCES

Cochran, W. G.(1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.

Ernst, L. R., and Ponikowski, C. H. (1995), “Specifications for Calculating Variance Components for an Integrated Compensation Survey,” Bureau of Labor Statistics memorandum to The Record, dated February 27.

Tehonica, J., Ernst, L. R., and Ponikowski, C. H. (1996), “Summary of Estimation and Variance Specifications for the 1996 Albuquerque, NM COMP2000 Test Survey,” Bureau of Labor Statistics memorandum to The Record, dated August 1.

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